EARLY DIAGNOSIS OF BREAST CANCER

Katja Hukkinen

Faculty of Medicine, University of Helsinki
Helsinki University Central Hospital
Department of Diagnostic Radiology
Helsinki, Finland

Academic dissertation

To be publicly discussed, with permission of
The Medical Faculty of the University of Helsinki,
in auditorium 1, Psychology department, Siltavuorenpenger 20 d
on November 30th 2007 at 12 noon.

Supervisor

Professor Leena Kivisaari
Department of Radiology
University of Helsinki

Reviewers

Professor Ritva Vanninen
Department of Radiology
University of Kuopio
and
Docent Tarja Rissanen
Department of Radiology
University of Oulu

Opponent

Professor Peter Dean
Department of Radiology
University of Turku

ISBN 978-952-92-2971-0 (paperback)
ISBN 978-952-10-4348-2 (PDF)
http://ethesis.helsinki.fi
Yliopistopaino, Helsinki 2007

To Jooa

ABSTRACT

The greatest effect on reducing mortality in breast cancer comes from the detection and treatment of invasive cancer when it is as small as possible. Accurate preoperative diagnosis of a breast lesion is essential for optimal treatment planning. In order to avoid unnecessary patient distress, it is important to achieve the definite diagnosis without delay and with as few biopsies as possible. Nowadays, when breast cancer is one of the most frequently diagnosed malignancies among women, cost-effective ways for its diagnosis are necessary.

The studies made aimed at evaluating whether breast cancers detected at screening had been visible in previous mammograms, and at assessing whether computer-aided detection (CAD) can help radiologists to detect these lesions. Aim was also to evaluate different reading methods and the impact of the number of readers on sensitivity and specificity. After detecting suspicious lesion, the radiologist has to decide what is the most accurate, fast, and cost-effective means of further work-up. For these reasons, the use of fine-needle aspiration cytology was compared to core-needle biopsy.

In the year 2001, 67 women with 69 surgically proven cancers detected at screening in the Mammography Centre of Helsinki University Hospital had previous mammograms as well. These mammograms were analyzed by an experienced screening radiologist, who found that 36 lesions were already visible in previous screening rounds. CAD (Second Look™ v. 4.01) found 23 of these "missed" lesions, and an inexperienced resident found 22; together they found 30 lesions.

The impact of CAD on reader's performance was evaluated next: Eight readers with different kinds of experience with mammography screening read the films of 200 women with and without CAD. These films included 35 of those "missed" lesions already mentioned and 16 screen-detected cancers. CAD sensitivity was 70.6% and specificity 15.8%. Use of CAD lengthened the mean time spent for readings but did not significantly affect readers' sensitivities or specificities.

With the same study setting (with only readings without CAD) two reading methods were compared in terms of sensitivity and specificity: "summarized" independent reading (at least a single cancer-positive opinion within the group considered decisive) and "conference consensus" reading (the cancer-positive opinion of the reader majority was considered decisive). The greatest sensitivity of 74.5% was achieved when the independent readings of 4 best-performing readers were summarized. Overall the summarized independent readings were more sensitive than "conference consensus" readings (64.7% vs. 43.1%) while there was far less difference in mean specificities (92.4% vs. 97.7%).

The feasibility of FNAC and CNB in the diagnosis of breast lesions was compared in non-randomised, retrospective study of 580 (503 malignant) breast lesions of 572 patients. Absolute sensitivity was 67% (194/289) for FNAC and 96% (206/214) for CNB ($p< 0.0001$), while complete sensitivities were 95% and 99%, respectively ($p=0.0173$). In patients with FNAC, an additional needle biopsy was performed for 93 and a surgical biopsy for 62 lesions. In the CNB group, a subsequent CNB was performed for 2 and a surgical biopsy for 33. The need for surgical biopsies and the unnecessary axillary operations due to false-positive findings raised the costs of FNAC from 150 to 294 € and 176 to 223 € for CNB, respectively.

Even though CAD is rather sensitive in finding "missed" cancer lesions, its impact on the readers' performance is minimal; it is therefore not recommendable to use this version of CAD device. The sensitivity of reading screening mammograms is maximal when any positive opinion within a pair or a group of readers is taken into consideration. The frequent need of supplement biopsies make FNAC more expensive than CNB, and because the advantage of quick analysis vanishes during the overall diagnostic and referral process, it is recommendable to use CNB as initial biopsy method.

LIST OF ORIGINAL PUBLICATIONS

This dissertation is based on the following papers, which are referred to in the text by the Roman numerals I-IV.

I) Hukkinen K, Pamilo M. Does computer-aided detection assist in early detection of breast cancer. Acta Radiologica 2005;46:135-139.

II) Hukkinen K, Pamilo M, Vehmas T, Kivisaari L. The effect of computer-aided detection (CAD) on mammographic performance. Experimental study on readers with different levels of experience. Acta Radiologica 2006;47:257-263.

III) Hukkinen K, Kivisaari L, Vehmas T. The impact of the number of readers on mammography interpretation. Acta Radiologica 2006;47:655-659.

IV) Hukkinen K, Kivisaari L, Heikkilä P, von Smitten K and Leidenius M. Unsuccessful preoperative biopsies, fine needle aspiration cytology (FNAC) or core needle biopsy (CNB), lead to increased costs in the diagnosis of breast cancer. Submitted to The Breast.

ABBREVIATIONS

ABBI = advanced breast biopsy instrumentation

ADH = atypical ductal hyperplasia

BI-RADS = breast imaging reporting and data system

CAD = computer aided detection

CL = confidence level

CNB = core needle biopsy

DCIS = ductal carcinoma in situ

FNAC = fine needle aspiration cytology

FN = false-negative

FP = false-positive

LCIS = lobular carcinoma in situ

MG = mammography

NHSBSP = National Health Service Breast Screening Programme

PPV = positive predictive value

TN = true-negative

TP = true-positive

US = ultrasound

VACB = vacuum-assisted biopsy

1. INTRODUCTION

One of the most frequently diagnosed malignancies among women is breast cancer. According to the Finnish Cancer Registry, it was the leading cause of cancer deaths among women in Finland in 2005. There were 4027 new breast cancer cases in 2005 (Finnish Cancer Registry). If breast cancer is detected early, more specific and less aggressive therapy options are possible, and mortality from breast cancer falls. Screening mammography has proved to be the most valuable single tool for reducing breast cancer mortality (Shapiro et al. 1982, Tabar et al. 1985, SOSSEG 1). It is, however, an unfortunate fact that observer errors in breast screening are frequent, and false-negative cancers emerge in retrospective studies of prior mammograms (Saarenmaa et al. 1999, Harvey et al.1993). Variability occurs in reading performances between mammography screeners and generalized radiologists as well as among experienced screeners (Sickles et al. 2002, Beam et al. 1996). To improve the quality of mammography screening, many screening centers use different types of double reading (Anttinen et al. 1993, Thurfjell et al.1994).

Computer-aided detection (CAD) systems have been developed to reduce the number of false-negative interpretations at screening mammography. Current CAD methods can achieve very high sensitivities of 71 to 98% in identifying malignant lesions, in which sensitivity is higher for microcalcifications than for tumor masses (Brem et al. 2001, Destounis et al. 2004, Moberg et al. 2001). The potential problem is the high number of false positives (Malich et al. 2001, Moberg et al. 2001, Thurfjell et al. 1998). Several studies examine readers' performance with CAD with different kinds of case selection and study design, most of them retrospective. Some have found with CAD an increase in the cancer detection rate with an insignificant rise in recall rate (Freer et al. 2001, Karssemeijer et al. 2003, Thurfjell et al. 1998), and especially that CAD helps more junior than senior radiologists (Balleyguier et al. 2005), while others found no significant change in radiologists' performance (Brem et al. 2001, Moberg et al. 2001, Taylor et al. 2004).

Correct preoperative diagnosis of a breast lesion is essential for optimal treatment planning. To reach as soon as possible the final diagnosis and operation with as few biopsies as possible is humane for the patient. Nowadays, when breast cancer is one of the most frequently diagnosed malignancies among women, cost-effective ways of diagnosis are crucial. Before the early 1990s, the recommended evaluation of a "suspicious" breast abnormality noted on either clinical examination or mammography involved a surgical breast biopsy. Nowadays, less invasive alternatives, fine-needle aspiration cytology (FNAC) and core needle biopsy (CNB), are useful in the evaluation of these breast abnormalities. Because of its high insufficient-sample rate and rather low specificity, FNAC has been increasingly often abandoned particularly in North America and the

UK (Britton et al. 1997, Pisano et al. 2001, Shannon et al. 2001). FNAC is, however, still considered a useful test in breast diagnosis, especially in assisting clinical decision-making whether to take additional biopsies or to proceed to surgical management (Bulgaresi et al. 2006).

2. REVIEW OF THE LITERATURE

2.1 Mammography Screening

Mass screening mammography program consists of invitation of asymptomatic women, mammography examination of breasts, reading of films, reporting the results to women and a systematic follow-up of the results and quality control. The primary purpose of screening mammography is to reduce mortality from breast cancer through early detection. Avoiding unnecessary workup of lesions which show clearly benign features will minimize anxiety and maintain streamlined, cost-effective service.

Mammography screening was first introduced in the United States in the 1960s. After that organized mammography screening has begun in many European countries as well, in England and Wales in 1988 (for women aged 50-64) (Blanks et al. 2000), in the Netherlands in 1989 (50-69) (Otto et al. 2003), in Sweden in 1980 (40-69) (SOSSEG I), and in Denmark in 1990 (50-69) (Olsen 2007). The European Union has recommended screening every 2 to 3 years women aged 50 to 69 (European Union Council 2003). In the United States, women over 40 are recommended for annual screening. In Finland, mammography screening started in 1987 and involved women aged 50 to 59. In some communities women aged 40 to 69 have also been screened. According to Section 14 of the Finnish Primary Health Care Act and the Government Degree on Screenings 1339/2006, as of 1 January 2007, screenings are also for women of 60 to 69. Screening takes place every second year, with two views, cranio-caudal and oblique, routinely taken. The technician fills in a special screening card, marking, for example, the palpation finding. The instructions of the Finnish Radiation and Nuclear Safety Authority (STUK), require consensus double reading. In the second and third screening rounds, the reader can compare films from previous rounds. If woman is recalled, the further studies include ultrasound (US), possible additional filming, and needle biopsies, if needed. During the first 11 years of screening in Finland, the compliance was 88.5%, recall rate 3.28% of those attending, and 0.65% were referred for surgery. Cancer-detection rate was 3.7 cancers per 1000 screening studies, and the specificity 97% (Dean and Pamilo 1999).

Prognostic indicators for breast cancer survival include tumor size, tumor type, tumor grade, mammographic presentation, regional lymph node status, and metastatic disease. The extent of disease evident from physical findings and special preoperative studies is used to determine its clinical stage. The American Joint Committee on Cancer and International Union Against Cancer have agreed on a TNM (Tumor, Regional Lymph Nodes, Distant Metastases) staging system for

breast cancer. Stages range from 0 to IV, in which the approximate 5-year survival is 95% for stage 0 and 5% for stage IV (Way 1994).

Invasive breast cancer must be detected at an early phase of its natural history (as small as possible), allowing the disease to be interrupted prior to the development of regional or systematic metastatic disease, in order to have the greatest effect on reducing breast cancer mortality. In a review article, Tabar and Dean (2003) stated that breast cancer is a progressive disease from the beginning. The progression can be arrested through early detection and treatment.

Many of the measures of a screening program involve statistical concepts like sensitivity and specificity.

Sensitivity = True Positive / (True Positive + False Negative), is a measure of the test's capability of finding, in this case, cancers in a population. Sensitivity decreases as false negatives increase. The goal is to maximize the detection rate of small invasive cancers.

Specificity = True Negative / (False Positive + True Negative) is a measure of how successful a test is at saying that cancer is not present when it really is not present. Specificity diminishes as false positives increase. False-positive interpretations and the further work-up after recall reduce the cost-effectiveness of screening mammography.

Accuracy = True Positives + True Negatives / number of all cases.

Recall Rate = number of women recalled for further work-up / number of women screened.

Cancer detection rate = number of cancers detected / 1000 screening studies.

Positive predictive value = True Positives / number of tests called positive for cancer. This maybe used to measure the discriminating power of the mammogram as a measure of what percentage is read as abnormal and needs additional imaging, or as a measure of the aggressivity of intervention when biopsies are recommended.

2.1.1 Efficacy of mammography screening

The Health Insurance Program of Greater New York (HIP) study, and later on The Breast Cancer Detection Demonstration Project (BCDDP) of 280 000 volunteer women, have been the pioneer projects investigating the feasibility of large-scale screening for breast cancer. HIP and BCDDP were the stimulus for development of the subsequent randomized clinical trials conducted in Europe and Canada (Cunningham 1997). The Edinburgh randomized trial, Swedish Two-County Trial and The updated overview of the Swedish randomized trials showed the significant impact of an invitation to mammography screening on mortality from breast cancer; reduction of 21-32% (Alexander et al. 1999, Tabar et al. 2000, Nyström et al. 2002). One Swedish study compared deaths from breast cancer diagnosed in the 20 years before introduction of screening (1958-77) with those from breast cancer diagnosed in the 20 years after the introduction of screening (1978-97). Taking account of potential biases, changes in clinical practice, and changes in breast cancer incidence, mammography screening contributed about 50% decrease in mortality when women actually attended the screening (Tabar 2003).

There has been debate over the effectiveness of mammography screening because of articles by Gotzsche and Olsen (2000 and 2001). They concluded that screening does not reduce mortality, and based their results on trials by Canadians. The Canadian studies compared annual screening with physical examination to physical examination only in women aged 40 to 49 and 50 to 59, and found that screening had no impact on breast cancer mortality (Miller et al. 2000 and 2002). Their results have been attributed to poor mammography technique and faulty trial design, which may have led to randomization errors.

One recent full cohort study showed that the estimated effect of routine mammography on breast cancer mortality to be highly dependent on study design; following the introduction of mammography screening the estimated changes in breast cancer mortality ranged from a 6% increase (a less rigid study design) to a 25% decrease (based on the best methodology) (Olsen et al. 2007).

After the success of randomized trials, organized service screening mammography has become routine in many countries. The results of these programs confirm that screening mammography reduces breast cancer mortality. The Swedish Organised Screening Evaluation Group (SOSSEG) derived results for 20 to 40 of observation on 1.1 million Swedish women, showing approximately 30% reduction in breast cancer mortality with invitation to screening and a 40% reduction for those receiving screening (SOSSEG 1 and 2). They also showed that screening had significantly (45% at 40-49 and 33% at 50-69) reduced the rates for larger tumors (>2 cm) and

the number of lymph-node- positive cancers in Sweden (SOSSEG 2007). Furthermore, Finnish studies of mammography service screening support these results (Hakama et al.1997, Anttila et al. 2002, Parvinen et al. 2006).

There is not consensus of opinion about screening women younger than 50 years, although the results are supporting. A randomized controlled trial in the United Kingdom with more than 160,000 women aged 39 to 48, showed a reduction in mortality, although not a significant one (Moss et al. 2006). Screening in Turku improved significantly breast cancer survival also at women aged 40 to 49 (Klemi et al. 2003).

2.1.2 Sensitivity of mammography screening

In a meta-analysis reported by Mushlin et al in 1998, the sensitivity of screening mammography ranged from 83 to 95%. A review article by Elmore et al. (2003) found a large international variation in screening mammography interpretation; North American screening programs appeared to interpret a higher percentage of mammograms as abnormal than did programs from other countries (by 2-4 percentage points) without any evident benefit in cancer detection rate (Elmore 2003). In a Finnish study, the sensitivity of mammography was significantly dependent on patient age and increased with age. Sensitivity was also dependent on the density of the breast, especially in the tumor area (Saarenmaa et al. 2001).

Digital mammography technique has been used in some screening centers, and according to first experiences, it seems to increase screening sensitivity for dense breasts and for younger women. In a study of 49 528 women in the United States and Canada, the overall diagnostic accuracy of digital and of film mammography as a means of screening for breast cancer, was similar. But the accuracy of digital mammography was significantly higher than film mammography among women under 50, women with heterogeneously dense or extremely dense breasts, and premenopausal or perimenopausal women (Pisano et al. 2005). The Oslo II Study found a higher cancer detection rate for full-field digital mammography in the age-group 50 to 69, and a nearly equal detection rate in the group aged 45 to 49 (Skaane et al. 2004). The overall reduction in average glandular X-ray dose for digital mammography was found over film mammography amounts to 27% (Gennaro et al. 2006). Digital image manipulation makes it possible to place images in a window, level them, and magnify them. Another advantage of digital mammography is the possibility of real-time interpretation of mammograms at distant sites with the use of teleradiology.

2.1.3 Classification of breast parenchymal patterns and findings on mammography

Breast parenchymal patterns can be classified in many ways: In the Tabar classification (Gram 1997): I = normal parenchymal pattern, II = fatty infiltration, III = retromamillary fibrosis, IV = adenosis and V = fibrosis. The Breast Imaging Reporting and Data System (BI-RADS) of the American College of Radiology is a tool created to standardize 1) terminology in mammographic reporting, 2) assessment of findings, and 3) resulting recommendation of action. BI-RADS lexicon describes four classes of breast parenchymal density and their effect on diagnostic accuracy (ACR 2003). The masses are classified in BI-RADS by shape, margin, and density. Irregular shape, ill-defined, obscured or spiculated margin and high density are definitions for a malignant mass lesion. Different types of calcifications and their distribution are also distinguished; grouped or clustered, granular and casting type of calcifications are considered to have a higher probability of malignancy. Architectural distortions can also be an associated finding of malignancy. BI-RADS categories range from 0 to 6, (0=needs additional imaging), 1 = negative, 2 = benign (their probability of malignancy is 0% and normal interval follow-up is recommended), 3 = probably benign (short-interval follow-up), 4 = suspicious abnormality (biopsy should be considered), 5 = highly suggestive of malignancy with probability of malignancy of $\geq$ 95% (appropriate action should be taken, and 6 = histologically proven malignancy (appropriate therapy recommended) (ACR 2003). BI-RADS lexicons are offered for ultrasound and magnetic resonance imaging findings as well.

2.1.4 Screening methods

Although mammography screening is known to be effective, observer errors are frequent and false-negative cancers can be found in retrospective studies of prior mammograms. The performance of the radiologist varies; those specialized in mammography may detect more early-stage cancers and have lower recall rates than do general radiologists (Sickles et al. 2002). Variability also exists in the detection of lesions among experienced radiologists, due to the complexity of breast tissue and the similarity between masses and normal tissue (Beam et al. 1996). Only a few breast cancers appear among a great number of normal and benign findings, and screener radiologists read hundreds of mammograms per day. This burden of work easily weakens their ability to detect malignant lesions.

To improve the quality of mammography screening, many screening centers use double reading. Objective independent double interpretation occurs when radiologists are unaware of each other's interpretations, and a standard rule is used to resolve differences. An average increase in sensitivity of 7% and a decrease in specificity of 11% occurred in the accuracy of a single reading versus an objective independent double interpretation with 31 community radiologists. This experimental test setting revealed no increase in accuracy (Taplin 2000). Independent prospective double reading occurs when the second reader is not blinded to results of the first reading (Destounis et al. 2004).

Consensus double reading occurs when two radiologists interpret independently all mammograms, and then all the cases selected by either reader are then reviewed by both readers, with a consensus decision made as to whom to recall for further studies. In a Finnish study, this method improved the breast cancer detection rate by 9% ($p<0.05$) and reduced the recall rate by 45% ($p<0.001$) (Anttinen et al. 1993). The cost of consensus double reading is competitive compared to single reading because it reduces the recall rate (Brown et al. 1996, Leivo et al. 1999).

In their review article, Dinnes et al. (10 cohort studies included) found that a double reading improves the cancer detection rate by 3 to 11 per 10 000 women screened and has a double impact on recall rates, depending on the recall policy used. This benefit was suggested to be mainly in the detection of small cancers, and could be greatest where two readers have different strengths and weaknesses, or where readers are less experienced (Dinnes 2001).

When two radiologists performing independent readings do not reach a consensus, an arbitration panel of three radiologists may be used to decide about referrals. Due to this kind of arbitration, 3% of cancers were missed in the study of Duijm et al. (2004). Arbitration by a third reader was the practice of Ciatto et al. (2005) when independent double reading produced the discordant reports which normally prompt diagnostic assessment. Their follow-up covered only 40% of subjects with negative arbitration and they observed 0.64% cancers as false-negative. Arbitration reduced recall rates by 32% and 49%, but due to the underlying low referral rates of both centers, the absolute reduction in referral rates was rather small, in the range of 1%. The cancers detected following arbitration are smaller and more likely to manifest as parenchymal distortion compared with cancers detected by both readers (Cornford et al. 2005).

Reader volume has been suggested as an important determinant of mammogram sensitivity and specificity. Esserman et al. (2002) had 194 high-volume U.K. radiologists and 60 U.S. radiologists run the same 60-film test. Radiologists were grouped as low-volume ($\leq$ 100 mammograms read per month), medium-volume (101-300 / month), and high-volume ($\geq$ 301 mammograms / month). The average sensitivity at specificity of 90% was 78.5% for U.K.

radiologists and 75.6% for high-volume U.S. radiologists, but only 64.8% for low-volume U.S. radiologists. The difference in sensitivity of these two groups of U.S. radiologists was statistically significant ($p > 0.001$) (Esserman 2002). The study of Kan et al. comes to the same conclusion; a minimum of 2 500 interpretations per year was associated with lower abnormal interpretation rates and average or better cancer detection rates (Kan et al. 2000). On the other hand, Beam et al. (2003) in their multifactor population study find no association between accuracy and reading volume. They suggested that "expertise reflects a complex multifactorial process; more recently trained radiologists interpreted mammograms more accurately than those trained earlier. The high number of diagnostic breast imaging examinations and image-guided breast interventional procedures increased accuracy". Moreover, use of double-reading and work in a dedicated mammography centre were associated with better accuracy. These findings of Beam et al. were in line with a more recent study by Barlow et al. (2004) who found no evidence that greater volume or even experience at interpreting mammograms was associated with better performance. They found that greater number of years of experience interpreting mammograms was associated with lower sensitivity ($p= 0.001$), but higher specificity ($p = 0.003$).

2.1.5 "Missed cancers"

Missed cancers are defined as those where biopsy-proved cancers are found on an asymptomatic subject's screening mammogram when the prior screening mammogram was prospectively interpreted as negative, but with the cancers judged retrospectively visible (Bird et al. 1992). When breast cancer is diagnosed mammographically, the corresponding lesion can be detected in the penultimate mammogram in up to 75% of cases (Bird et al. 1992, Harvey et al. 1993). Mass lesions were the commonest feature of missed lesions (80 of 115), and 32 of those 80 were spiculated or irregular in the 2001 study of Birdwell et al. The most frequently recorded factors related to lesion interpretation errors were lucent areas within the mass and benign-appearing calcifications.

Subtle findings define non-specific findings perceptible in an initial negative screening mammogram but that subsequently develop into cancer. These kinds of findings are often seen in clinical practice, but do not necessarily warrant a recall (Ikeda et al. 2003). In the multicenter study of Warren Burhenne et al. (2000) 67% (286 of 427) of lesions were visible in prior mammograms, but only 27% (115 of 427) were interpreted by a panel of radiologists as warranting a recall.

Interval cancers are defined as cancers in patients presenting with clinical findings before the next scheduled mammogram. True-interval cancers are not visible in a prior mammogram, and 42

to 89% of interval cancers have been found to be those (Evans et al. 2007, Simpson et al. 1995, Frisell et al. 1987, Ikeda et al. 1992, Saarenmaa et al. 1999). Blanks et al. found that at least 40% of false negative interval cancers corresponded to lesions originally detected and misclassified as benign (Blanks et al.1999).

Most of these retrospectively visible cancer lesions occurred in dense breasts, and were asymmetric densities or architectural distortions (Bird et al. 1992, Harvey et al. 1993, Ikeda et al.1992, Saarenmaa et al. 2001). In younger patients the most frequently missed or misinterpreted feature were granular microcalcifications (38%) (Evans et al. 2007).

2.2 Computer-aided detection (CAD)

Computer-aided detection (CAD) is aimed towards reducing the false-negative rate of screening mammography by marking suspicious mass lesions and microcalcifications. CAD is not designed to replace the radiologist, but rather to help the radiologist in detecting possible abnormal lesions. In 1998, the US Food and Drug Administration approved the first CAD system for clinical use in screening mammography. Today, the Food and Drug Administration has approved three CAD systems for clinical use: ImageChecker (R2 technology, Sunnyvale, CA), Second Look (CADx Medical Systems, Laval, Quebeq, Canada) and MammoReader (Intelligent Systems Software, Clearwater, FL). In the use of film-screen mammography, the CAD system needs a film digitizer, a processing computer, and an image display system but in the case of digital mammography, the CAD system may be integrated. The CAD system makes special marks at the location of microcalcifications, masses, and dense areas.

2.2.1 Sensitivity of CAD

Sensitivity of CAD systems in retrospective studies ranged from 76 to 98% (screen-detected cancers), with sensitivity being higher for microcalcification than for masses. There were 0.5 to 1.3 false-positive marks per image (Morton et al. 2006, Brem et al. 2005, Malich et al. 2001, Karssemeijer et al. 2003). The threshold for abnormality detection is adjustable, so when the sensitivity of the system is high, then the number of false-positives also increases. In a study by Hoffmeister et al. (2002) ImageChecker and Second Look showed no significant differences in sensitivities. In a recent study, R2 ImageChecker M1000, version 5.0A was significantly more

sensitive for finding small (<16 mm), noncalcified mass lesions than iCAD Second Look, version 6.0, 81.8% vs. 60.9% (Ellis et al. 2007).

The sensitivity of the CAD system deteriorated significantly as the density of the breast increased, whereas the specificity remained relatively constant (Ho et al. 2003). Brem et al. (2005) and Malich et al. (2005) found that breast density does not impact overall CAD detection of breast cancer, but the sensitivity for mass lesions in dense breasts may be lowered.

The sensitivity for finding "missed" lesions has ranged from 62 to 86%, with again sensitivity being higher for microcalcifications (Karssemeijer et al. 2003, Destounis et al. 2004, Warren Burhenne et al. 2000, Birdwell et al. 2001). CAD is reported to have the potential to reduce the false-negative rate at a double reading by more than one-third (from 31 to 19%) (Destounis et al. 2004).

2.2.2 Experimental studies

In experimental study settings of readings with and without the CAD system have had various impacts on readers' performance. In the studies by Thurfjell et al. (1998) (51 cancers; one screening and one clinical radiologist; readings in different sessions) and Balleyguir et al. (2005) (13 cancer; one senior and one junior reader; readings at the same session), they found that CAD might be more helpful for less experienced mammography readers than for expert screeners. Taylor et al. (2004) compared the performance with and without CAD with 30 radiologists, 5 "breast clinicians", and 15 radiographers (60 cancers; readings in different sessions) and found no difference in impact of CAD for readers performing well or poorly. Nor did they find any significant change in sensitivity or in specificity, but they complained about the small number of difficult cancers (16) regarding which real improvements could be detected. They then carried out a new study using a data set of 40 cancers missed by at least one radiologist at screening but which were marked by CAD; the sensitivity improved but was slightly below the threshold for statistical significance (Taylor et al. 2004). Ciatto et al. (2003) reported a significant increase in sensitivity as well as a significant decrease in specificity when 19 readers viewed 11 false-negative cases and 20 with minimal signs. Taplin et al. (2006) found increased specificity but no affect on sensitivity from use of CAD. Breast density did not seem to affect CAD's performance. In the study of 10 radiologist reading mammograms showing clustered calcifications, variation in accuracy was reduced by 46% with CAD aid (Jiang et al. 2001).

2.2.3 Prospective studies

When the same images in clinical settings were interpreted first without and then with CAD, the increase in cancer detection ranged from 4.7 to 19.5% (Ko et al. 2006, Morton et al. 2006, Freer et al. 2001, Helvie et al. 2004). Freer et al. had two radiologists who read 12 860 screening mammograms first without and subsequently with CAD aid in the same sessions. They reported a 19.5% improvement in cancer detection rates. One limitation may be that they had only two film readers, who knew while viewing the mammograms without CAD that they would have the "safety net" of CAD in the next reading. In the prospective study by Ko et al., 5 016 mammograms (48 cancers) were interpreted with and without CAD in a manner suggested by developers: the reader could not change the recall decision to a negative assessment based on lack of CAD marking. Two additional cancers emerged in a 26-month study period, in a 12-month follow-up, three interval cancers developed; one of which was marked by CAD but had been ignored by the reader. The follow-up was based on the hospital's cancer registry and therefore was incomplete; there might have been more false-negatives. For those 2 additional cancers found, an additional 89 women were recalled, and an additional 6 were biopsied. The increase in absolute recall rates were 1.2 to 2.0%.

When historical changes were compared before (56 432 mammograms) and after (59 139 mammograms) installation of the CAD system, Gur et al. (2004) reported no significant increase (1.7%) in their cancer detection rate. Cupples et al. (2005) reported 16.1% of increase in cancer detection rate but due to sample-size limitation their differences did not reach significance (7 872 mammograms read without CAD and 19 402 with CAD). They found also a younger age at diagnosis and a significantly earlier stage of invasive cancer detection. Recall rates remained unchanged 11.4% (Gur) or increased from 7.7 to 8.3% (Cupples).

2.3 Ultrasound and magnetic resonance imaging

Even though mammography is the best way for breast cancer screening, additional modalities are needed for further evaluation and diagnosis of breast lesion. Ultrasound (US) plays an important role as an examination complementary to mammography, especially in younger women and those with dense breast parenchyma. Axillary areas are imaged with US in order to find suspicious lymph nodes, and needle biopsies of breast lesions are most often taken under US-control. Breast US can detect additional cancers and also downgrade lesions from cancer-positive to cancer-negative when used with mammography (Saarenmaa et al. 2001, Flobbe et al. 2003). Malignant features in US are

irregular shape, microlobulated or spiculated margins, and reduction or shadowing of the beam. Benign lesions are usually round or oval in shape, with circumscribed margins and enhancement of the beam.

Breast MRI is most useful for preoperative evaluation of breast carcinoma, and for screening asymptomatic BRCA mutation carriers. Traditionally MRI has been regarded to be more sensitive than mammography in imaging of invasive cancers but according to new results it is more sensitive for in situ cancers as well. In a prospective study of 7 319 women, 167 women with DCIS preoperatively had both MRI and mammography: MRI was more sensitive in detecting these lesions 92% versus 56%. MRI was especially helpful in diagnosis of high nuclear grade tumors. (Kuhl et al 2007). Women genetically predisposed to breast cancer often develop the disease at a young age when dense breasts reduce the sensitivity of X-ray mammography. Contrast-enhanced MR imaging was significantly more sensitive than mammography, 77% versus 40% (Leach et al 2005).

In a prospective study of 111 women with known cancer, in nonfatty breasts, US and MR imaging were more sensitive than mammography for detection of local extent of invasive cancer. Combined mammography, clinical examination, and MR imaging were more sensitive than any other individual test or combination of tests (Berg et al. 2004).

2.4 Breast lesion biopsy

Breast lesion biopsies include surgical biopsies, or less invasive needle biopsies. Biopsies have been taken under palpation control, in current practise imaging guidance is recommended; US guidance, or if lesion is invisible in ultra-sound, it can be localized from mammography by use of a 3-dimensional stereotactic technique.

2.4.1 Fine needle aspiration cytology (FNAC)

A 20- to 25-gauge needle with a 10-ml syringe in a special holder is used to aspirate cells from the lesion. Usually, two to five separate aspirations are made from different areas of the lesions.

A pathologist performs the cytological analysis. In Finland the cytological grading of Papanicolaou is used: 0 = insufficient, 1 = normal, 2 = probably benign, 3 = mildly suspicious of malignancy, 4 = strongly suspicious of malignancy, 5 = malignant. The National Health Service Breast Screening Program (NHSBSP) has proposed another classification for cytological results,

which is nowadays widely used in Europe and United States: C1, inadequate sampling, C2, benign lesion, C3, probably benign lesion, C4, suspicious of malignancy, and C5, presence of carcinoma-like cells.

In one study of US-guided FNAC, 213 malignant and 580 benign, palpable, and non-palpable lesions were aspirated. C4 and C5 combined gave a sensitivity of 95%. Specificity was 92% and accuracy 93%, when 27 (3%) C1 specimens were included in the group of C2 (Gordon 1993). US-guided FNAC resulted in 13.6% fewer insufficient aspirates and 16.5% better sensitivity (C4-5) than did freehand FNAC in a study of 102 palpable breast cancers (Houssami et al. 2005). Sensitivity of 97.5% (suspicious and positive outcomes combined) was found versus 90% sensitivity of CNB in the Ballo et al. (1996) study of 124 palpable breast lesions (mean size, 4.4 cm), but CNB did contribute to a more definitive diagnosis in some cases.

In his review article, Britton (1999) introduced terms “absolute sensitivity” and “complete sensitivity”, better way for comparing results from several studies; absolute sensitivity means unequivocal malignancy (Papa 5, C5), and complete sensitivity means categories of atypia, suspicious of malignancy, and malignancy (C3-5, Papa 3-5). Absolute sensitivities for FNAC were 62.4% (stereotactic biopsy) and 83.1% (ultrasound biopsy), complete sensitivities were 83.1% and 95.1%. Specificities for these procedures were 86.9% and 84% (Britton 1999).

Advantages of FNAC include rapidity of the procedure and a possibility for immediate diagnosis, easy performance in outpatient settings, relative painlessness without need for local anesthesia, low cost, easy mastery of the technique, and easy clinicopathological correlation.

Disadvantages of FNAC include relative high rate of insufficient samples, 8 to 34% (Britton 1999, Pisano 2001, Lieske 2006), being higher for calcifications (46%) and for other nonpalpable lesions in stereotactic guidance (40%) (Pisano 1998). The false-negative rate has been reported as 2 to 13% (Ciatto 1992, Lieske 2006). The false-negative results are partially the result of underreporting of abnormalities that may be noted at review (Bulgaresi 2005). A major disadvantage is inability to distinguish invasive from in situ lesions; FNAC showed an overestimation for DCIS of 9% (Pijnappel 2004).

2.4.2. Core needle biopsy (CNB)

14G needles are preferred in automated guns compared to 16G and 18G needles (Nath et al. 1995). For the exact tumor classification, Parker et al. (1994) suggested a standardized protocol approach that includes a prone biopsy table, long-throw (2.3 cm excursion) biopsy gun, 14G cutting needles

and removal of five or more tissue samples per lesion. In their study, CNBs were taken with both US and stereotactic guidance (Parker 1994). The average specimen weight with a 14-gauge automated biopsy was 18 mg in a study with a turkey-breast parenchymal model (Berg et al. 1997. A Finnish study found that more than three samples are needed for a histologic diagnosis of a mass lesion by use of an add-on stereotactic biopsy device (Koskela et al. 2005). With three-dimensional ultrasound, three CNB were found to be sufficient for reliable histological diagnosis (Sauer et al. 2005).

In Finland, pathologists give a specific descriptive statement of the outcome of CNB, and the following classification according to European guidelines is used: normal/insufficient for diagnosis of lesion or calcification, specific benign diagnosis (e. g. fibroadeoma, fat necrosis), benign with minor risk of malignancy (e. g. atypical ductal hyperplasia (ADH), radial scar), suspicious for malignancy, or malignant (European guidelines for quality assurance in mammography screening-fourth edition). The classification of CNB outcomes is proposed by NHSBSP in the following way: B1, unsatisfactory/normal tissue only; B2, benign; B3, benign with minor or uncertain malignant potential; B4, suspicious of malignancy; B5, malignant (B5a non-invasive cancer, B5b invasive cancer, B5c cancer of nonassessable invasiveness). Absolute sensitivity as suggested by Britton includes only B5 and complete sensitivity B3-B5.

A multi-institutional study showed follow-up (surgical follow-up for 1 363 lesions) results of 3 765 lesions with 14G needle biopsy; complete agreement between needle biopsy and surgical biopsy was reached in 1 223 (89.7%) lesions, partial agreement in 125 (9.2%) and disagreement (false-negatives) in 15 (1.1%) lesions (Parker 1994). In their review, absolute sensitivities for stereotactic biopsy and ultrasound biopsy were 90.5% and 96.7%, and complete sensitivities were a respective 94.6% and 98.5%. Specificities for CNB were 98.3% for stereotactic and 98.7% for US-biopsy (Britton 1999). In a study by Jackman et al. (1999), stereotactic CNB was performed with a 14-gauge needle for 483 nonpalpable lesions in order to estimate false-negatives and underestimation rates. Some lesions were confirmed by another biopsy or in surgery, or lesions were mammographically followed-up. The false-negative rate was 1.2%, 58% of ADH were carcinoma in surgery, and the histologic underestimation rate for DCIS was 15% (Jackman 1999). In a prospective multicenter COBRA study (COre Biopsy after RAdiological localisation), 1 029 nonpalpable lesions were offered streotactic 14G needle biopsy. Of the 871 successful biopsy procedures, 95% were confirmed surgically. There was 1.5% with insufficient material and 4% false-negatives; 23% of the high-risk lesions were diagnosed as cancers in surgery (Verkooijen et al. 2002).

Equal absolute sensitivity (75% vs. 72%), positive predictive value (PPV) for malignancy (99% vs. 100%), and inadequate rate (7% vs. 7%) was found for US-guided CNB and FNAC taken from the same lesions (n = 286) and by the same operator. But the specificity for CNB was significantly better (90 vs. 82%), differences emerged for PPV of suspicious and atypia categories, and for the suspicious rate. In all of these cases CNB performed better (Westenend et al. 2001). Another study of 869 malignant lesions (both CNB and FNAC taken from the same lesion) found absolute sensitivity of 80% for CNB and 65% for FNAC. The same trend appeared regardless of guidance method (Lieske et al. 2006). Conversion to CNB from FNAC in a single center increased absolute and complete specificity, and reduced inadequate rate and suspicious rates in preoperative diagnosis of both the screen-detected and symptomatic practice (Shannon et al. 2001). Together FNAC and CNB can lead to an increase in absolute sensitivity without affecting specificity, and a decrease in the inadequate rate for cancers (Westenend et al. 2001, Lieske et al. 2006).

Advantages of CNB include low rate of insufficient samples (Britton 1999), a better definition of atypical and malignant lesions (Sun et al. 2000), and minimum errors in interpretation. CNB also offers an option for the grading and typing of tumors and assessment of progesterone and estrogen receptor status and HER-2 by immunocytochemistry, thereby making diagnostic information more available for treatment options (Shannon et al. 2001, Al Sarakbi et al. 2005, Sutela et al. 2007).

Disadvantages of CNB include need for local anesthesia, it is time-consuming and expensive, and there is a risk for complications such as hematomas (Sun et al. 2000). Discussion continues as to the increased risk of local recurrence after core-needle biopsy on patients treated with breast conservation surgery and radiation therapy. Recent studies have suggested otherwise (Chen et al. 2002, Fitzal et al. 2006).

2.4.3. Vacuum-assisted biopsy

Vacuum-assisted biopsies (VACB) are taken with 8-, 11-, and 14-gauge probes (e. g. Mammotome, Biopsys/Ethicon Endo-Surgery, Cincinnati, OH, USA and Minimally Invasive Breast Biopsy, United States Surgical, Norwalk, CT, USA). This probe can be inserted once, with multiple specimens (5 to 64) obtained from a single insertion. The tissue is acquired at a distance from the probe and the volume is larger than with a 14-gauge core-needle biopsy (37 mg for 14-gauge and 94 mg for 11-gauge) (Berg 1997). The vacuum-assisted biopsy probe enables placement of a localizing clip to mark the biopsy site. VACB can be performed under US- or stereotactical guidance, and the procedure requires local anesthesia. In the review article by Plantade et al. (2004) 2 130 VACB

were evaluated to assess its reliability in diagnosing atypical ductal hyperplasia and DCIS. The resection revealed an underestimation of 10 of 37 (27%) for ADH and of 12 of 319 (3.8%) for DCIS. Three-dimensional US-guided and stereotactical VACB allows even complete excision of benign breast lesions and thus avoidance of open surgery and postoperative scarring (Pfarl et al. 2002, Baez et al. 2003).

The largest series of automated (14G) and directional vacuum-assisted (11G) CNB reported thus far is a study by Ciatto et al. (2007), reporting the accuracy of 4 035 core biopsies. The overall yield of malignancy was 30%. The overall sensitivity of CNB was 94.2% (92.9-95.5%) and specificity 88.1% (86.6-89.6%). Positive predictive value was 84.8% (82.9-86.7%) and negative predictive value was 95.6% (94.6-96.6%). The insufficiency rate was 0.57%, which decreased with greater operator caseload; the false-negative rate was 4.4% (3.4-5.4%). Vacuum-assisted biopsy (11G) was more sensitive than 14G automated core but their overall accuracy was very similar. The overall underestimation of CNB was 27.7%, with a significant trend toward greater underestimation of malignancy with increasing lesion size on imaging. A larger needle gauge and vacuum-assisted core devices reduced underestimation as did increasing the total volume of tissue obtained at CNB (Houssami et al. 2007a). From the same series, of the 372 lesions with uncertain malignant potential (borderline histology) identified, approximately one-third were malignant on excision. Atypical ductal hyperplasia and lobular intraepithelial neoplasia were associated with higher probability of cancer, whereas phyllodes tumor and radial scar had the lowest PPV for malignancy (Houssami et al. 2007b).

2.4.4. Cost-effectiveness of different biopsy techniques

Utilization of FNAC and CNB (14-gauge) is a cost-effective way to spare women from unnecessary open surgical biopsy without compromising breast carcinoma detection (Logan-Young et al. 1998). The article by Liberman and Kaplan (2001) reviewed studies comparing the utility and costs of percutaneous core biopsies of nonpalpable lesions to surgical biopsies. Ultrasound-guided core biopsy spared the patient a surgical procedure in 85% of cases, reducing the cost of diagnosis by 56%, while the same numbers were 76 to 85% and 40 to 58% with stereotactic 14-gauge automated core biopsy, and respectively 76% and 20% with stereotactic 11-gauge VACB.

Liberman et al. (2001) suggested using stereotactic 11-gauge VACB for calcifications highly suggestive of malignancy rather than CNB to achieve the highest cost savings (Liberman et al. 2001).

Altomare et al. compared the use of FNAC and Advanced Breast Biopsy Instrumentation (ABBI) or vacuum biopsy (VACB) in sequence instead of surgical biopsy. The surgical procedure was thus avoided in 97.1% cases, yielding a 73.9% decrease in the cost of diagnosis compared with that for surgical biopsy (Altomare 2005).

3. AIMS OF THE STUDY

I: To evaluate whether breast cancers detected at screening had been visible in previous mammograms and to assess a computer-aided detection system in detecting lesions in preoperative and prior mammograms.

II: To evaluate the effect of computer-aided detection on readers' performance.

III: To study the effect of different numbers of readers on sensitivity and specificity of mammography, and to compare independent reading of several radiologists to conference consensus reading.

IV: To compare feasibility of FNAC and CNB in diagnosis of breast lesions. The special aim was to evaluate the extra costs and delay in surgical treatment due to unsuccessful preoperative biopsies.

4. MATERIALS AND METHODS

All the study protocols were approved by the Ethics Committee of Helsinki University Hospital.

4.1 Material

Mammography films of a total of 200 women, used in Studies I, II and III, were filmed in mammography screening in Helsinki during the years 1997 to 2001. In 2001, 104 women were referred for surgery, and 83 breast cancers were detected in 81 women. Of those 104 women operated on, 23 exhibited benign histological findings. Of the 81 women operated on for breast cancer, 67 had mammograms taken in earlier screening rounds. Of these 67 women, 65 had breast cancer in 1 breast and 2 in both breasts. The preoperative mammograms and mammograms from previous screening rounds of these 67 women with 69 cancers were used in Study I.

Studies II and III, involved previous films of 33 women with 35 cancers, preoperative films of 16 cancer patients, and 8 women with operated benign findings. In addition to those films, another 125 women had mammograms that showed no malignancy in 4 years' follow-up.

In Study IV, the material included a consecutive 572 women with 580 lesions operated on at the Breast Surgery Unit of Helsinki University Central Hospital between 1 February 2005 and 30 January 2006. The postoperative diagnosis was ductal carcinoma in situ (DCIS) or invasive cancer for 503 lesions in 496 patients. Patients operated on because of aberrant breast tissue and hypertrophic breasts, as well as those with suspicion of local recurrence after mastectomy, were excluded. Furthermore, 12 patients with vacuum-assisted biopsy as well as 50 patients with CNB obtained under stereotactical guidance were excluded, because these biopsy methods are used on occasions when FNAC is not recommended, such as for diffuse microcalcifications. In 575 lesions, a radiologist performed the initial percutaneus biopsies under US guidance or under palpation. On two occasions, the biopsies, one FNAC and one CNB, were obtained under palpation check by a surgeon. Two FNAC and one CNB were taken from microcalcifications under a coordinate grid technique. FNAC was the first biopsy method for 339 lesions and CNB for 241. Most of the private clinics and small radiology divisions have not introduced CNB and have taken only FNAC, whereas the larger breast-imaging units have used primarily CNB. In these units FNAC is used if the location of the lesions is not safe for CNB or it is not technically possible to take CNB. In the CNB group, the lesions on imaging were larger; more lesions had a mass combined with microcalcifications, fewer lesions were invisible in a mammogram, and more lesions had invasive

lobular cancer as the postoperative diagnosis. Otherwise, between the FNAC and CNB groups, lesion characteristics were well balanced (Table 1.)

Table 1. Characteristics of the 580 breast lesions undergoing fine needle aspiration cytology (FNAC) or core needle biopsy (CNB)

	FNAC (n=339)		**CNB (n=241)**		**p-value**
	n	%	n	%	
Clinical feature					
Median patient age		59y		58y	
Symptomatic	203	60	137	56	0.0528
Palpable	237	70	166	69	0.8549
Median size in imaging		15 mm		20 mm	
Size < 10 mm	75	22	29	12	0.005
Range of size in imaging		3-80 mm		5-60 mm	
Finding in mammogram					
Mass	245	72	169	70	0.5774
Mass+ microcalcification	17	5	34	14	0.0002
Architectural distortion	20	6	20	8	0.31
Microcalcification	9	3	9	4	0.47
Invisible	29	9	6	2	0.0023
Only US taken	19	6	3	1	0.007
Finding in ultrasound					
Benign in US	32	9	20	8	0.6614
Intermediate in US	39	12	20	8	0.2647
Malignant in US	266	78	200	83	0.2034
Invisible in US	2	0.6	1	0.4	1
Suspicious lymph nodes	24	7	10	4	0.15
Postoperative diagnosis					
Benign	50	15	27	11	0.2637
DCIS	8	2	9	4	0.4545
IDC	194	57	124	51	0.1763
ILC	59	17	65	27	0.0074
Other invasive cancers	28	8	16	7	0.5265

Abbreviations: DCIS = ductal carcinoma in situ, IDC = invasive ductal carcinoma, ILC = invasive lobular carcinoma. Fisher's exact test for the two-tailed p-values was used.

For 274 lesions, breast imaging and biopsies were performed or directly supervised by experienced radiologists at the Department of Radiology of HUCH. Of these biopsies, 80 (29%) were FNAC and 194 (71%) CNB. For 293 lesions, the biopsies were performed by radiologists at private clinics; 250 (85%) were FNAC and 43 (15%) CNB. The remaining 12 specimens, 8 FNAC and 4 CNB, were obtained in public health care hospitals other than HUCH. In HUCH, the CNB were taken

with a 14-gauge core needle. Information regarding needle gauge was not available from all of the referral units.

The cyto- or histopathological assessment of 281 percutaneus biopsies, 86 FNAC and 195 CNB was performed or directly supervised by experienced breast pathologists at the Department of Pathology of HUCH. The remaining 299 specimens, 253 FNAC and 46 CNB, were assessed at other public health care- or private pathology laboratories by specialist pathologists. All the surgical breast specimens were assessed at the Department of Pathology of HUCH.

4.2. Methods

4.2.1. Finding the "missed lesions"

Of the 81 women with breast cancer, 67 had mammograms taken in earlier screening rounds. These previous films of 67 women with 69 breast cancers were read first without knowledge of the preoperative mammograms. Lesions in previous mammograms that an experienced radiologist thought suspicious for cancer were compared to cancer lesions in preoperative mammograms. Those suspicious lesions in prior mammograms that matched the cancer lesions in preoperative mammograms were considered "missed" lesions. For Study II, these "missed lesions" were classified "actionable" if they had distinct features suggesting breast cancer, or "subtle finding" if the appearance of the lesion was less specific.

The imaging findings in mammography were described as stellate/spiculated mass, circumscribed mass or only a mass (Study IV), microcalcification, mass + microcalcification or architectural distorsion. Sizes of the lesions on imaging were also noted.

Breast parenchymal structure in mammograms was classified into five groups based on the pattern classification introduced by Tabar (Gram 1997). Lesion histology and lymph node involvement was noted from TNM staging (Tumor, Regional Lymph Nodes, Distant Metastasis) (Study I).

4.2.2. CAD

The films used in Studies I (all the mammograms from previous screening rounds and preoperative mammograms), and II and III (all mammograms of 200 women) were analyzed by Second Look™

(v. 4.01; CAD$_X$Systems, Inc., Beavercreek, OH, USA). The system has a film digitizer with a 43.5μ-pixel resolution with 12 bits per pixel (Howtek). The system assigns microcalcifications and tumor masses with special marks (CalcMark™ and MassMark™). The Paper print Mammagram™ was used to analyze the results. Every woman had two views taken from each breast: craniocaudal and mediolateral oblique. A correct mark in one projection was treated as a true-positive. If two marks overlapped, they were treated as one. All CAD-prompted sites in prior mammogram were checked from preoperative mammograms. If CAD made marks in cancer patient's healthy breast or marked incorrect locations in a cancer breast, these were considered false-positives, whereas a healthy breast without marks by CAD was treated as true-negative.

4.2.3. Methods used in study I

Study I compared size and morphology of lesions in prior and preoperative mammograms as well as the performance of CAD on detecting these lesions. A resident with one year experience in mammography read previous mammograms without seeing preoperative mammograms. The sensitivities and specificities of CAD and resident were compared.

Sensitivity = True Positive/ (True Positive + False Negative)
Specificity = True Negative/ (False Positive + True Negative)

No statistical comparisons were made in Study I.

4.2.4. Methods used in Studies II and III

From Study I, films of 33 women, with 35 "missed" cancers, were available at the time of Study II. In addition to those cancer lesions on prior mammograms, the material included 16 screen-detected and histologically proven cancers, 8 operated benign lesions, 25 non-operated benign lesions, and 316 breasts without lesions (in 4 years' follow-up they showed no malignancy). The material was divided onto two rollers, films of 100 women in each roller: eight screen-detected cancers on each roller and the retrospectively found actionable or subtle cancers randomly divided between rollers.

Test readers from the Helsinki area were invited, and eight radiologists volunteered. This test group comprised four radiologists with experience in screening mammography ranging from 5 to 18 years, two general radiologists, and two residents, each of whom had 6 months to 4 years of

experience in clinical mammography. None of the readers had prior experience in CAD. They were given an explanation of prompting and of the behavior of the CAD system, thus becoming aware of the rather high sensitivity and low specificity of CAD. They were told that the material was enriched with breast cancers but were unaware of the number of cancers. Readers worked independently. The readers received numbered breast maps, on which they drew the area and feature of suspected pathological lesions: stellate mass, circular mass, microcalcification, or architectural distortion. The correct feature in the correct location was considered a true positive. Readers were asked not to mark benign lesions.

Roller A contained the films of 100 women; 19 of 26 cancers (73%) were marked by CAD. In Study II, the readers viewed films first without CAD and then, after a minimum interval of 7 days, again with CAD. For both sessions they could spend as much time as necessary. Roller B also contained films of 100 women, 17 of 25 cancers (68%) were marked by CAD. Readers also viewed this roller twice: first without CAD, and again after 7-day minimum interval, with CAD. Only for roller B did they have a time limit: 60 minutes for the screening radiologists and 90 minutes for other readers. They recorded the time they used in each of the four sessions. For both rollers, the time interval between readings without and with CAD ranged from 7 to 21 days (mean 12 days).

Every film reader's reading with and without CAD was cross-tabulated with the pathological finding as a layer factor, and the sensitivities and specificities of these two methods were then compared by use of the modified McNemar test (Hawass 1997) and with a small sample size formula introduced by Conover (1980) was used. Making multiple observations per woman (both breasts) may induce bias in p-values. Statistical tests were therefore adjusted for this bias by taking the intracluster correlation into account as suggested by Gönen et al. in 2001. The Yates' correction was used for continuity. The total numbers of correct and incorrect diagnoses for rollers A and B were compared with the chi-square test.

Every change between readings without and with CAD was compared to CAD prompts to see whether the changes were resulted from CAD. The sensitivity, specificity, and false-positive rate of CAD were then calculated.

The time used for image reading was analyzed with the general linear model. The independent variables (use of CAD vs. no CAD, free reading schedule vs. prompted reading schedule, and the reader effect) were all simultaneously entered into the model. The first two were fixed factors and the third a random factor. SPSS 12.0 software (SPSS Inc. Chicago, IL, USA) served for the analysis.

Study III used only the readings without CAD. Each breast (n = 400) was classified as cancer-positive or cancer-negative by each radiologist. The radiologists' readings were cross-tabulated with

the pathological findings (positive or negative). The sensitivities, specificities, and accuracies were calculated. The readers were arranged in order from 1 to 8 according to their sensitivities. The independent double-reading of each possible reading pair (28) was calculated.

The independent readings of virtual groups of different sizes were assessed by accepting even a single positive opinion as the group decision. This was done by considering groups consisting of the best readers 1 and 2; 1, 2 and 3; 1, 2, 3, and 4, and so forth until all 8 were included in the group. The same was done by starting from the least sensitive readers 8 and 7; 8, 7 and 6 and so on.

Conference consensus readings were defined as the majority opinion in the group; in case of a paired number of readers and equal votes the cancer-positive opinion prevailed (for instance, 3/6 readers). The calculations were performed in the same manner as described above.

4.2.5. Methods in study IV

The data were collected retrospectively from the patient records. The type of initial and potential subsequent biopsies with their results were recorded and also compared with the histological findings of the surgical specimen. Cytological grading of Papanicolau (0 = insufficient, 1 = normal, 2 = probably benign, 3 = mildly suspicious of malignancy, 4 = strongly suspicious of malignancy, 5 = malignant) was used for FNAC. CNB findings were classified as insufficient, benign, indeterminate, suspicious of malignancy, or malignant. The indeterminate category consisted of findings that warranted a surgical biopsy: atypical ductal hyperplasia (ADH), lobular carcinoma in situ (LCIS), papilloma, radial scar, or benign tumor phyllodes.

Absolute sensitivities - number of invasive or in situ cancers diagnosed as Papa 5 or invasive/in situ cancers in CNB divided by the total number of invasive or in situ cancer - were calculated, and complete sensitivities - number of invasive or in situ cancers diagnosed as Papa 3 to 5 or indeterminate to cancer in CNB divided by the total number of invasive or in situ cancers. Differences between sensitivities of FNAC and CNB groups were calculated by Fisher's exact test for the two-tailed P-values and by Newcombe's method for 95% confidence levels.

Time interval was calculated between initial biopsy and the definite surgical treatment among 490 women with in situ or invasive breast cancers. Patients who received neoadjuvant therapy (n = 4) or those with the operation postponed due to comorbidity (n = 2) were excluded from this analysis.

When biopsy result was insufficient for treatment planning and needed additional work-up, or a false biopsy result caused unnecessary operations, these extra expenses were recorded. For

patients with benign, indeterminate, or insufficient findings in the needle biopsy but invasive cancer in the surgical specimen, axillary surgery was performed as a second operation. In these cases, the price of the surgical biopsy was added. In patients with benign or indeterminate findings in the needle biopsy but DCIS with insufficient resection margins in the surgical biopsy, no extra charges were added. If a patient was operated on because of an indeterminate biopsy finding, and the postoperative diagnosis was not malignant, the expenses of the surgical biopsy were charged. When the initial or additional CNB showed DCIS, but invasion was detectable in the surgical specimen, the expenses of axillary surgery were not charged when SNB was included in the operation. In cases with unnecessary axillary surgery because of a false-positive biopsy finding, the costs of axillary surgery were added, but not in cases with FNAC PAPA 4 to 5 and DCIS for the surgical specimen.

Because of the variety of referring units, the HUCH price-list for radiological, pathological, and surgical interventions was used (Table 2).

Table 2. Prices at Helsinki University Central Hospital

ITEM	**€**
Radiology	
US-guided CNB	136
US-guided FNAC	95
Pathology	
Histology of CNB	40
Cytology of FNAC	55
Histology of surgical biopsy specimen	100
Histology of sentinel node	365
Surgery	
Inpatient care per day	300
Outpatient visit	75
Surgical biopsy under local anesthesia	680
Surgical biopsy under general anesthesia	1120
Sentinel node biopsy	1520

Abbreviations: US = ultrasound, CNB = core needle biopsy, FNAC = fine needle aspiration cytology.

5. RESULTS

5.1. Study I

In 2001, 67 women with previous mammograms, had 69 breast cancers surgically removed (Group A). A radiologist experienced in screening mammography detected 36 (52%) lesions visible in mammograms from previous screening rounds (Group B). Of these 36 lesions, 27 were visible in mammograms taken 2 years earlier (in 1999); seven lesions were visible in both 1997 and 1999 mammograms. Four years earlier (in 1997), two patients underwent mammograms which already showed signs of the lesion. The proportion of small mass lesions was higher in previous mammograms (Group B, "missed" lesions) than in preoperative mammograms (Group A, preoperative lesions) (Table 3).

Table 3. Characteristics of lesions visible in mammograms 2 to 4 years before diagnosis ("missed" lesions) and lesions on preoperative mammograms in 2001

	Missed lesions (n=36)		**Preoperative lesions** (n=69)	
	n	%	n	%
Number of women	34		67	
Visible in 1997	9	25		
Visible in 1999	27	75		
Finding in mammogram				
Stellate	22	61	37	54
Circumscribed	7	19	12	17
Microcalcification	3	8	13	19
Architectural distortion	4	11	7	10
Size < 10 mm	14	39	6	9
Size 10-20 mm	20	55	50	72
Size > 20 mm	2	6	13	19
Postoperative diagnosis				
Invasive carcinoma	34	94	66	96
In situ cancer	2	6	3	4

The distribution of lesions histology and breast parenchymal patterns was similar in Groups A and B. Half the cancers were ductal, and between 22 and 29 % were lobular carcinomas. Those

exhibiting a normal parenchymal pattern (I) amounted to 40%, 35% exhibited adenosis (IV), and the rest showed evenly fatty infiltration (II), retromamillary fibrosis (III), and fibrosis (V).

At the time of operation, the breast carcinoma had spread to ipsilateral lymph nodes (N1) in 25 (37%) of the 67 patients. Of these 25, 13 exhibited lesions visible in previous mammograms (Group B).

Mammograms from 2001 (Group A) were analyzed with CAD, which correctly marked 56 (81%) of the 69 breast cancers (Table 4.)

Table 4. True-positive marks by CAD in preoperative mammograms: Group A (n:69)

Mammographic feature	**Size in mm**	**TP**
Stellate	<10	5/5
	10-20	20/28
	>20	4/4
Circumscribed	<10	1/1
	10-20	9/10
	>20	1/1
Microcalcification	<10	-
	10-20	4/7
	>20	6/6
Architectural distortion	<10	-
	10-20	4/5
	>20	2/2

Of 65 healthy breasts CAD registered 63 (97%). Only two healthy breasts scored true-negatives. The sensitivity of CAD to accurately register breast cancer in Group A was 81%; the specificity of CAD was only 3%, and CAD marked a mean of 1.2 false-positives per film.

Previous mammograms (films of 67 women) were analyzed with CAD, which correctly marked 23 (64%) of the 36 visible malignant lesions (Table 5). CAD prompts were compared to cancer lesions in preoperative mammograms, and CAD did not find additional “missed” lesions compared to those found by an experienced radiologist.

Table 5. True-positive marks by CAD in previous mammograms performed a mean 2 years before surgical treatment: Group B (n = 36)

Mammographic feature	Size in mm	TP
Stellate	<10	5/9
	10-20	8/12
	>20	1/1
Circumscribed	<10	2/4
	10-20	1/3
	>20	-
Microcalcification	<10	1/1
	10-20	1/2
	>20	-
Architectural distortion	<10	-
	10-20	3/3
	>20	1/1

Of 23 cases, CAD marked lesions accurately in 10 in both projections: one 8-mm stellate lesion, five 10-to 20-mm stellate lesions, one 5-mm and one 10-mm cluster of microcalcifications, one 8-mm circumscribed lesion and one architectural distortion 15-mm in diameter. The lesion of architectural distortion was already visible in 1997 and 1999. In both years, CAD marked the lesion but then missed it in the preoperative mammogram in 2001. A similar result was found with the 8-mm stellate lesion, which was correctly marked in both projections in the previous mammogram in 1999 but produced a false-negative in the preoperative mammogram in 2001. CAD missed six breast cancers, sizes 8 to 13 mm, in both previous and preoperative mammograms. Four of these lesions were stellate, one was circumscribed, and one consisted of microcalcifications. Overall, the sensitivity of CAD was 64% in Group B. Of 32 healthy breasts, CAD incorrectly registered 29, meaning that only three healthy breasts were correctly designed as true-negatives. Altogether, CAD registered between one and seven false-positives per woman. As a result, the specificity of CAD was low (9%).

The resident found 22 of the 36 malignant “missed lesions” in Group B, but made 11 false-positive findings, resulting in a sensitivity of 61% and a specificity of 89%.

Of the 22 lesions detected by the resident, CAD marked 15 in addition to detecting eight other “missed lesions.” The distribution of those eight lesions’ histology, morphology, size, and breast parenchymal pattern resembled that of all 36 lesions in Group B. Together, of 36 lesions in Group B CAD and the resident found 30 (83%).

Of these 67 women, 33 had normal previous mammograms. The breast parenchymal pattern distribution was essentially the same as in breasts with “missed” lesions. CAD marked 1.34 false-

positives per film (5.4 marks per woman). None of the marks occurred in locations where carcinomas later appeared, and as many false-positive masses appeared as did false-positive microcalcifications. The results from all previous mammograms together produced a mean of 1.1 false-positives per film. Of these 134 healthy breasts, CAD assigned no false-positives to only 3.

5.2. Study II

Of 51 cancers, CAD correctly marked 36, 22 of them in both projections. For roller A, sensitivity was 73% (19/26), and for roller B, 68% (17/25) (Tables 6 and 7).

Table 6. Cancer material in roller A

	Operated	Actionable	Subtle
Breast parenchyma			
Dense	8	8	8
Fatty	-	2	-
Carcinoma			
In situ	2	1	1
Invasive	6	9	7
Mammographic features			
Mass	6	7	5
Calcification	2	1	1
Mass + calcification	-	2	2
Architectural distortion	-	-	-
Tumor size			
Mean size mm	22	9	9
Range mm	9-60	6-20	5-15
CAD-positive	5/8	8/10	6/8

The dense breast category includes breast parenchyma patterns I, IV, V; the fatty-breast category includes patterns II and III according to the Tabar classification. Operated = screen-detected cancers, Actionable = feature suggesting breast cancer visible in previous mammogram, Subtle = non-specific feature visible in previous mammogram.

Table 7. Cancer material in roller B

	Operated	Actionable	Subtle
Breast parenchyma			
Dense	4	13	1
Fatty	4	2	1
Carcinoma			
In situ	-	1	-
Invasive	8	14	2
Mammographic features			
Mass	7	11	1
Calcification	-	1	-
Mass + calcification	1	3	-
Architectural distortion	-	-	1
Tumor size			
Mean size mm	11	9	21
Range mm	8-15	7-13	20-22
CAD-positive	7/8	8/15	2/2

Explanation as in Table 6

Specificity of CAD was 16% (28 true-negative of 174 control breasts) for roller A and 15% (27 true-negative of 175 control breasts) for roller B. Approximately 3.2 false-positive prompts occurred per cancer patient and 3.5 false-positives per non-cancer woman. Among 349 healthy breasts were 55 breasts with true-negative findings by CAD.

The time used for the image reading was significantly dependent on observer ($p < 0.001$), on reading schedule (mean time for unlimited time schedule 65 min. vs. 55 min. for reading with time limit, $p = 0.013$), and marginally dependent on use of CAD (no CAD 56 min. vs. 63 min. with CAD, $p = 0.053$). Screening radiologists were faster in reading both rollers without or with CAD (Table 8).

Table 8. Times of readings in minutes. No time-limit for roller A. For roller B, time-limits: 60 minutes for screeners, 90 minutes for general radiologists

	A without	A with CAD	B without	B with CAD
Screener 1	45	65	45	60
Screener 2	60	73	60	60
Screener 3	50	50	30	45
Screener 4	40	45	30	35
Radiologist 2	70	70	80	80
Resident 3	120	120	90	85
Resident 4	40	50	30	40

Radiologist 1 did not provide information.

As radiologist 1 did not record reading time, calculations include the reading times of only 7 readers.

Mean sensitivities were 42% for screeners vs. 45% for others reading roller A without CAD, and were 42% vs. 48%, respectively, with CAD. For roller B, with a time schedule, mean sensitivities were 38% (screeners) vs. 36% (novice radiologists and residents) without CAD, and 42% vs. 29% with CAD (Tables 9 and 10).

Table 9. Sensitivity and specificity of readings without and with CAD. Unlimited time-schedule (Roller A). Numbers of true-positive (TP) and false-positive (FP) findings in parentheses. (n = 26 cancer breasts; total 200)

	Sensitivity	**Sensitivity**	**Specificity**	**Specificity**
Screener 1	23.1 (6)	23.1 (6)	97.1 (5)	97.1 (5)
Screener 2	30.8 (8)	42.3 (11)	96.0 (7)	95.4 (8)
Screener 3	65.4 (17)	57.7 (15)	94.3 (10)	94.8 (9)
Screener 4	46.2 (12)	46.2 (12)	96.6 (6)	97.1 (5)
Radiologist 1	34.6 (9)	34.6 (9)	98.3 (3)	98.3 (3)
Radiologist 2	46.2 (12)	50.0 (13)	95.4 (8)	98.3 (3)
Resident 3	61.5 (16)	73.1 (19)	95.4 (8)	93.7 (11)
Resident 4	38.5 (10)	34.6 (9)	98.9 (2)	97.7 (4)

P-values for differences between individual sensitivities and between individual specificities were notsignificant ($p > 0.05$). P-values were derived from the modified McNemar test, the small sample formula was used, and intracluster correlation was taken into account.

Table 10. Sensitivity and specificity of readings without and with CAD. Reading times: 60 minutes for screeners, 90 for others (Roller B). Numbers of true-positive (TP) and false-positive (FP) findings in parentheses. (n = 25 cancer breasts; total 200)

	Sensitivity	**Sensitivity**	**Specificity**	**Specificity**
Screener 1	28.0 (7)	36.0 (9)	99.4 (1)	99.4 (1)
Screener 2	28.0 (7)	36.0 (9)	98.9 (2)	97.7 (4)
Screener 3	48.0 (12)	52.0 (13)	96.0 (7)	93.7 (11)
Screener 4	48.0 (12)	44.0 (11)	96.6 (6)	97.1 (5)
Radiologist	20.0 (5)	20.0 (5)	98.9 (2)	98.9 (2)
Radiologist	40.0 (10)	28.0 (7)	96.0 (7)	97.7 (4)
Resident 3	52.0 (13)	48.0 (12)	96.6 (6)	94.9 (9)
Resident 4	32.0 (8)	20.0 (5)	98.9 (2)	97.7 (4)

See Table 9 for explanations

Differences between individual sensitivities or between individual specificities in readings without and with CAD were not significant (p-values >0.05). When readings of roller A and B were combined, the sensitivities for screeners ranged from 26 to 57% without CAD and from 29 to 55%

with CAD. For other readers the combined sensitivities ranged from 28 to 57% without and from 28 to 61% with CAD. The best sensitivity, 61%, was reached by one resident in readings with CAD (vs. 57% without CAD). A screening radiologist had the poorest combined sensitivity, 26%, reached in readings without CAD, but the radiologist's sensitivity rose to 29% with CAD.

In 3200 readings, CAD caused 62 changes (Table 11). Of 408 cancer readings, 17 cancers were found with CAD aid: 3 times a screen-detected cancer, 13 times an "actionable" cancer, and once a "subtle" cancer. On roller A both groups found six new cancers with the CAD aid. On roller B the screeners found four additional cancers, and radiologists made only a single new cancer diagnosis. No significant difference appeared in the number of correct and incorrect diagnoses between rollers A and B. CAD missed 15 cancers, and because of the false-negatives, readers' true-positive findings changed to false-negative in 10 readings. Screeners changed their true-positive finding to false-negative four times because of CAD but only on roller A. Radiologists changed true-positive findings to false-negatives three times on both rollers. Roller A had four cancers (three of them positive by CAD) that none of the readers found, and roller B, nine similar unidentified cancers (six of them positive by CAD). CAD made false-positive markings on 294 breasts (2352 readings), and readers made 25 new false-positive diagnoses because of CAD (1% increase).

A change in readings occured 36 times, contrary to CAD's prompting (Table 11).

Table 11. Changes in readings caused by CAD (n=62) and despite CAD markings (n = 36). Rollers A and B combined

CAD	**TP**		**FN**		**TN**		**FP**	
	FN→TP	**TP→FN**	**TP→FN**	**FN→TP**	**FP→TN**	**TN→FP**	**TN→FP**	**FP→TN**
Screener 1	2	0	0	0	0	0	0	0
Screener 2	3	0	0	2	1	0	6	2
Screener 3	3	3	2	1	2	2	5	2
Screener 4	2	2	2	1	1	0	1	2
Radiologist	0	0	0	0	0	0	0	0
Radiologist	3	3	2	0	5	0	2	6
Resident 3	3	0	2	1	1	1	9	3
Resident 4	1	3	2	0	0	2	2	0
Total	17	11	10	5	10	5	25	15

TP = true-positive, FN = false-negative, TN = true-negative, FP = false-positive

5.3. Study III

The two most sensitive readers showed same sensitivity of 56.9%, so their difference in specificities governed their ranking order. All readers had specificities more than 95%, but the sensitivities as low as 25.5% (Table 12).

Table 12. Sensitivities, specificities and accuracies of each reader. N = 51 cancerous and 349 non-cancerous breasts. (TP = true-positive and FP = false-positive in parentheses).

	Sensitivity % (TP)	Specificity % (FP)	Accuracy%
Resident 1	56.9 (29)	96.0 (14)	91.0
Screener 2	56.9 (29)	95.1 (17)	90.3
Screener 3	47.1 (24)	96.6 (12)	90.3
Radiologist 4	43.1 (22)	95.4 (16)	88.8
Screener 5	35.3 (18)	98.9 (4)	90.8
Resident 6	29.4 (15)	97.4 (9)	88.8
Radiologist 7	27.5 (14)	98.6 (5)	90.0
Screener 8	25.5 (13)	98.3 (6)	89.0

Only one reader (a radiologist) found all 16 screen-detected cancers, the rest found 10 to 15 screen-detected cancers. The false-negative cancers were more challenging: of the 35 false-negative cancers, the best performing reader (a resident) found 15, while the rest of the readers found 3 to 14. None of the readers found exactly the same cancers. Of the 28 possible pairs of readers, only 3 showed no increase in the number of true-positives compared to the true-positives of the better reader in the pair (reading pairs 2 and 8, 1 and 6, 4 and 6). The specificities of these independent double readings dropped 0.3-1.3% from the specificity of the better reader. The best combination was the pair of screening radiologists 2 and 5; their independent double reading showed a sensitivity of 67% and specificity of 94%.

The combined sensitivity of the two best-performing readers was 67%, as compared with 57% for the single best reader. The specificity was 93% vs. 96% and accuracy 90% vs. 91%. The greatest sensitivity of 75% was attained after summarizing the readings of the four best readers (with the positive opinion of at least a single reader considered decisive); 9 additional cancers were detected, as compared to the 29 cancers found by the best-performing reader. For every additional true-positive finding, two false-positive cases emerged, leading to specificity decreasing from 96 to 91% (Table 13).

Table 13. Sensitivities, specificities and accuracies of independent and conference readings of different numbers of readers. N = 51 cancerous and 349 non-cancerous breasts. (TP = true-positive and FP = false-positive in parentheses). Calculations starting with the most sensitive readers.

	Independent			Conference		
Readers	**Sensitivity % (TP)**	**Specificity % (FP)**	**Accuracy %**	**Sensitivity % (TP)**	**Specificity % (FP)**	**Accuracy %**
1+2	66.7 (34)	93.4 (23)	90.0	66.7 (34)	93.4 (23)	90.0
1+2+3	70.6 (36)	92.8 (25)	90.0	52.9 (27)	96.6 (12)	91.0
1+2+3+4	74.5 (38)	91.1 (31)	89.0	54.9 (28)	96.3 (13)	91.0
1+2+3+4+5	74.5 (38)	91.1 (31)	89.0	47.1 (24)	97.4 (9)	91.0
1+2+3+4+5+6	74.5 (38)	90.0 (35)	88.0	47.1 (24)	97.4 (9)	91.0
1+2+3+4+5+6+7	74.5 (38)	90.0 (35)	88.0	43.1 (22)	98.6 (5)	91.5
All 8	74.5 (38)	90.0 (35)	88.0	45.1 (23)	99.1 (3)	91.3

Independent readings = the positive opinion of at least a single reader considered decisive. Conference consensus readings = the positive opinion of the readers' majority considered decisive, in case of a paired number of readers and equal votes, the cancer-positive opinion prevailed.

Starting the calculation from the least sensitive reader, the sensitivity improved until cancer-positive findings of all eight readers were summarized. The conference consensus readings, on the contrary, decreased in sensitivity while increasing in specificity (Table 14).

Table 14. Sensitivities, specificities and accuracies of independent and conference readings of different number of readers. N = 51 cancerous and 349 non-cancerous breasts. (TP = true-positive and FP = false-positive in parentheses). Calculations starting from the less sensitive readers.

	Independent			Conference		
Readers	**Sensitivity % (TP)**	**Specificity % (FP)**	**Accuracy %**	**Sensitivity % (TP)**	**Specificity % (FP)**	**Accuracy %**
8+7	39.2 (20)	97.1 (10)	89.8	39.2 (20)	97.1 (10)	89.8
8+7+6	45.1 (23)	96.3 (13)	89.8	23.5 (12)	98.3 (6)	88.8
8+7+6+5	51.0 (26)	96.0 (14)	90.3	33.3 (17)	98.3 (6)	90.0
8+7+6+5+4	56.9 (29)	93.1 (24)	88.5	31.4 (16)	98.9 (4)	90.3
8+7+6+5+4+3	58.8 (30)	92.0 (28)	87.8	39.2 (20)	98.6 (5)	91.0
8+7+6+5+4+3+2	70.6 (36)	90.8 (32)	88.3	35.3 (18)	98.9 (4)	90.8
All 8	74.5 (38)	90.0 (35)	88.0	45.1 (23)	99.1 (3)	91.3

See notes in Table 13.

The mean sensitivity for summarized independent readings of different groups was 65% (33 TP) compared to 43% (22 TP) mean sensitivity of conference consensus readings. The mean specificities were 92% (26.5 FP) and 98% (8 FP), respectively.

5.4. Study IV

The absolute sensitivity was 67% (194 of 289) for FNAC and 96% (206 of 214) for CNB ($p < 0.0001$, difference of 29%; 95% CL 22.9-35.0). The complete sensitivity of FNAC was 95% (273 of 289) and 99% (211 of 214) for CNB ($p = 0.0173$, difference of 4%; 95% CL 0.8-7.5). The absolute sensitivity of the CNB was superior to that of FNAC regardless of tumor characteristics, but these findings did not reach significance for lesions that were few (Table 15). Of 339, FNAC resulted in 48 (14%) Papa 3; of these 48 lesions, 36 were malignant. Of 241, CNB revealed 9 (4%) indeterminate findings (ADH, LCIS, papilloma, radial scar or benign tumor phyllodes); the final diagnosis was malignant for five.

Table 15. Absolute sensitivity of fine needle aspiration cytology (FNAC) or core needle biopsy (CNB) in diagnosis of 503 malignant breast lesions.

	FNAC % (n= 289)	**CNB % (n=214)**	**p-value**	**Difference% (95 % CL)**
Palpability				
Palpable	74 (157/213)	97 (145/149)	0.0001	24 (16.8-30.1)
Non-palpable	50 (38/76)	94 (61/65)	0.0001	44 (29.9-55.4)
Size in imaging				
< 10 mm	45 (24/53)	93 (26/28)	0.0001	48 (27.2-61.2)
> 10 mm	73 (171/236)	97 (180/186)	0.0001	24 (17.9-30.6)
Ultrasound				
No finding	50 (1/2)	0 (0/1)	1	
Benign	33 (3/9)	50 (2/4)	1	
Indeterminate	47 (9/19)	73 (8/11)	0.2595	25 (10.6-52)
Suspicious	70 (182/259)	99 (202/204)	0.0001	29 (22.9-34.6)
Mammogram				
No finding	54 (15/28)	100 (4/4)	0.12	46 (5.4-64.2)
Microcalcification	57 (4/7)	86 (6/7)	0.55	29 (17.3-62.7)
Architectural distortion	50 (8/16)	100 (20/20)	0.0004	50 (22.7-72)
mass	72 (159/222)	96 (146/152)	0.0001	24 (17.4-31)
Mass+ microcalcification	56 (9/16)	97 (30/31)	0.0012	41 (16.1-63.7)
Histology				
DCIS	25 (2/8)	78 (7/9)	0.0567	53 (5.7-76.7)
IDC	74 (143/194)	98 (121/124)	0.0001	24 (16.6-30.7)
ILC	64 (38/59)	97 (63/65)	0.0001	33 (19.2-45.5)
Other invasive cancers	43 (12/28)	94 (15/16)	0.001	51 (22.4-68)

Abbreviations: DCIS = ductal carcinoma in situ, IDC = invasive ductal carcinoma, ILC = invasive lobular carcinoma, CL = Confidence level. Differences between sensitivities of FNAC and CNB groups were calculated by Fisher's exact test for the two-tailed P-values and by Newcombe's method for 95% confidence levels.

No significant difference (p = 0.0921) emerged between the sensitivity of FNAC obtained at HUCH, of 62, 36 (52%) malignant lesions, and at other private or public health care units: of 227, 159 (70%) malignant lesions.

Of the 580 operated lesions, 297 were spiculated masses in mammography and suspicious for malignancy in US. Only one of these 297 had a benign postoperative histology. It was correctly diagnosed on CNB as papilloma. Invasive carcinoma was the final postoperative diagnosis for 294 of the other lesions, with the remaining two being DCIS. FNAC was performed on 177 of these lesions and CNB on 119. The sensitivity of the CNB group was 98%, but for FNAC only 75% ($p < 0.0001$, difference of 24%; 95% CL 17-31% (Table 16).

Table 16. Results of US-guided fine needle aspiration cytology (FNAC) (n=177) and US-guided core needle biopsy (CNB) (n=119) in the 296 lesions with malignant postoperative histology which appeared as spiculated masses on mammography and suspicious for malignancy on ultrasound.

Biopsy/Lesion size in imaging	FNAC Papa 0	FNAC Papa 1	FNAC Papa 2	FNAC Papa 3	FNAC Papa 4	FNAC Papa 5
FNAC < 10 mm (n=25)	1 (4%)	0	0	4(16%)	5 (20%)	15
FNAC ≥ 10 mm (n=152)	1 (1%)	3 (2%)	4 (3%)	9 (6%)	18(12%)	117(77%
	CNB	**CNB**				
CNB < 10mm (n=20)	1 (5%)	19				
CNB ≥ 10mm (n=99)	1 (1%)	98				

Of the 339 lesions with FNAC as the initial biopsy method the finding in FNAC was malignant, Papa 5, in 197. An additional CNB was performed for four lesions, and cancer surgery was planned for 193; in these 193 lesions, invasive cancer was detected in 190, DCIS in one, and a benign finding in two of the surgical specimens.

In the remaining 142 lesions with Papa 0-4 for the initial biopsy, a subsequent FNAC was obtained for 4 lesions and a subsequent CNB for 85. In addition, for a total of 62 lesions a surgical biopsy was performed, revealing invasive cancer in 11 lesions and DCIS in 4. Cancer surgery without further biopsies was performed in three cases with benign or indeterminate findings in the subsequent CNB and definitely malignant findings in imaging; invasive cancer was detected in the surgical specimen in all three cases. Cancer surgery was performed without additional biopsies for 19 lesions: The postoperative diagnosis was invasive cancer in 16, DCIS in 2 and benign in one (Figure 1).

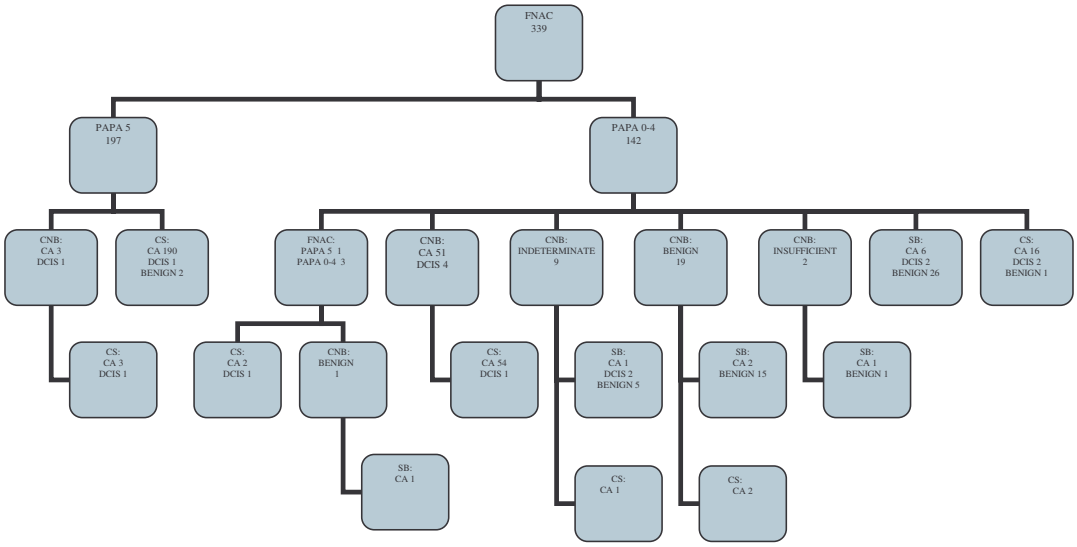

Figure 1. Diagnostic work-up in the 339 breast lesions with fine needle aspiration cytology (FNAC) as the initial biopsy. Abbreviations: CNB = core needle biopsy, SB = surgical biopsy, CS = cancer surgery, CA = invasive cancer, DCIS = ductal carcinoma in situ.

In the CNB group, a subsequent CNB was necessary for 2 lesions. A surgical biopsy was performed for 33 lesions, revealing 4 invasive cancers and 2 DCIS. In addition, invasion was detected in the surgical specimen in 2 lesions with DCIS in the initial CNB. Cancer surgery without further biopsies was performed on one patient with an indeterminate finding in the initial CNB and a definitely malignant finding in imaging. The surgical specimen revealed invasive cancer (Figure 2).

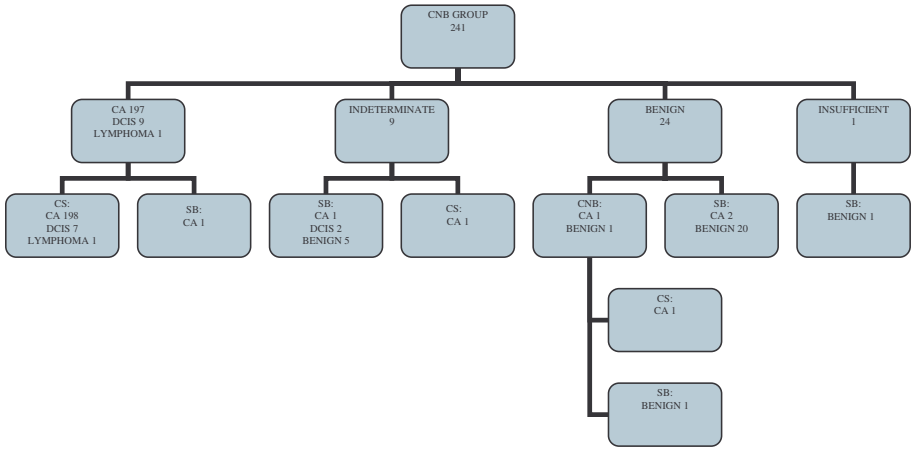

Figure 2. Diagnostic work-up in the 241 breast lesions with core needle biopsy (CNB) as the initial biopsy. Abbreviations as in Figure 1.

In both groups, when the biopsy showed cancer, the median time from initial biopsy to definite surgical treatment was 31 days. The extra needle and surgical biopsies delayed the definite surgical treatment for 75 of the 283 patients with FNAC but for only 5 of the 207 patients with CNB: 27% vs. 2%. All cancer patients with FNAC as their initial biopsy method had their operation a mean of 3 days later than did patients with CNB. In addition, one patient with bilateral cancer had FNAC for the one and CNB for the other breast. Both the FNAC and the CNB failed to detect cancer, and surgical cancer treatment was delayed accordingly (Table 17). Six patients with neoadjuvant treatment or co-morbidity postponing surgery were excluded from this analysis.

Table 17. Time interval in days between first biopsy and definite surgical treatment for 490 women with in situ or invasive breast cancers with fine needle aspiration cytology (FNAC) or core needle biopsy (CNB).

Biopsy	Median days	Mean days	Range of days
FNAC all (n=283)	32	35	6-101
Papa 5 (n= 193)	31	32	6-66
Papa 0-4 (n=90)	39	42	10-90
Without additional biopsies (n= 208)	31	31	6-101
With additional needle biopsy (n=69)	42	44	13-90
With surgical biopsy (n=6)	63	62	27-85
With additional needle and surgical biopsy (n=	67	64	39-85
CNB all (n=207)	31	32	6-77
Cancer (n=200)*	31	31	6-64
Insufficient/benign/indeterminate (n=7)	41	43	13-77
Without additional biopsies (n= 202)	31	31	6-64
With additional needle biopsy (n=1)	13	13	13
With surgical biopsy (n=4)*	51	50	35-65

* Includes one lymphoma

In the FNAC group, the extra expenses resulted from 93 supplemental needle biopsies, 11 surgical biopsies of malignant lesions, and 8 surgical biopsies of benign lesions. Three patients underwent unnecessary axillary surgery, and these costs were charged. One of these three women had 3 extra inpatient care days because of a wound infection; the costs of these extra days were also included.

In the CNB group, the extra expenses were due to two supplemental needle biopsies, four surgical biopsies of malignant lesions, and five surgical biopsies of benign lesions plus one unnecessary sentinel node biopsy performed in a patient with a postoperative diagnosis of lymphoma instead of ductal carcinoma.

The cost of the initial biopsy was 150 € per lesion for FNAC and 176 € per lesion for CNB. With the expenses caused by the additional needle biopsies included, corresponding costs were 210 € for FNAC and 177 € per lesion in the CNB group. The need for surgical biopsies and the unnecessary axillary operations due to false-positive findings raised the costs to 294 € in lesions with FNAC as the initial biopsy and to 223 € in lesions with CNB.

If the CNB had been initial biopsy method instead of FNAC, the saving would have been 24%.

99 666 € (339 FNAC x 294 €) – 75 597 € (339 CNB x 223 €) = 24 069 €.

6. DISCUSSION

Breast cancer is the leading cause of cancer deaths in women in developed as well as in developing countries. The incidence of breast cancer in Finland in 2005 was 87.1 per 100 000 women according to the Finnish Cancer Registry. This study focuses on early diagnosis of breast cancer.

6.1. Missed lesions

Mammography screening is a demanding task; despite of double reading many cancers are missed at screening. As many as half the cancers found in mammography screening in our center during the year 2001 were to some degree visible in previous mammograms 2 to 4 years earlier. Small mass was the most common feature of "actionable" missed lesions found, in line with other findings (Birdwell et al. 2001, Ikeda et al. 2003, Karssemeijer et al. 2003). Some of these lesions were slowly growing tumors, but some were more aggressive and at the time of diagnosis had already spread to axillary lymph nodes.

Some may criticize the way the missed lesions were found; retrospectively with the suspicion that some of those patients developed cancer later on. Possibly lesions found this way were non-specific, subtle, and non-actionable and too difficult for CAD to detect. However, most of these lesions were detected by the resident (Study I), and later on also by test readers (Studies II and III).

6.2. CAD performance

CAD showed some potential: it marked correctly 64% of those malignant lesions missed in consensus double reading. Mass lesions within dense breast tissue were more difficult for CAD to detect, in line with other studies (Brem et al. 2005, Malich et al. 2005). The sensitivity of CAD to detect preoperative lesions was better, 81%, showing that missed lesions, most of them mass lesions, were difficult also for CAD. Alberdi et al (2005) found a strong correlation between reader and CAD errors; cases that were difficult for human readers to interpret were also difficult for CAD. This is easy to understand because the algorithms of computer-aided detection devices are programmed by human.

Because the resident and CAD found different "missed" lesions it was assumed that CAD might have the potential to a help reader to find these lesions, especially if the reader were less

experienced with mammograms. However, because the high number of false-positives raised doubt as to CAD's effectiveness in practice, therefore in Study 2, readers with different levels of experience of mammograms read films without and with the aid of CAD.

6.3. Effect of CAD on reading

Difficult cases, missed lesions among preoperative lesions, in test settings helped to elucidate the potential utility of CAD. The sensitivity of CAD (70.5%) and the number of false-positives (average 3.2 per cancer patient and 3.5 per non-cancer woman) corresponded to others' findings (Birdwell et al. 2001, Moberg et al. 2001, Karssemeijer et al. 2003).

The cancer cases used were challenging also for test readers, as shown by low sensitivities, the lowest sensitivity without CAD being 25% (13 of 51). One possible reason to this is that fear of missing cancers may have not stressed the readers. Specificities on the other hand were high, over 95% (for both rollers) in spite of the fact that there was no economic pressure of high recall rates during the test. This artificial test situation may have had a different kind of influence on different readers. Especially the non-screeners may have found mammography reading without clinical examination and US strange. Competence in mammography interpretation and also the threshold for marking suspicious lesions as cancers may have been very different among the readers. Longer experience may even raise the threshold for calling a mammogram positive and therefore lead to a low sensitivity but high specificity, as was evident in the study by Barlow et al (2004).

The changes in sensitivities between readings with and without CAD were statistically nonsignificant. The maximal changes in sensitivities were 12% to better or to worse (±3 cancers) ($p>0.05$). High specificities also remained at the same level, with no significant changes found.

Ciatto et al. (2003), with 31 interval cancers and 19 readers, found a significant rise in sensitivity with CAD aid. This may have been due to the fact that they assumed that all true-positive findings in conventional single reading were automatically positive also by CAD reading. Their argument for this was that CAD is not intended to alter the radiologist's recall decision if CAD does not mark an area that a radiologist has detected on the initial mammography evaluation (Ciatto et al. 2003).

The change in reader's sensitivity reflects not only the effect of CAD on the reader's decision-making. The reader may find new cancers or miss cancers that he/she has already found in the unprompted situation despite the CAD markings. All the individual changes between readings were studied, whether they were made according to CAD markings or contrary to CAD advice.

A positive effect, in which a reader found new cancer with CAD help, occurred in 17 of 408 "cancer readings." When the four CAD-positive cases that all readers detected even without CAD and the nine CAD-positive cases that no one found were excluded, there remained 23 cancers that CAD marked correctly. Why did some readers then ignore these lesions that were sufficiently visible and pointed out by CAD? Performance of CAD may explain part of this phenomenon; because the threshold of CAD was adjusted to maximum sensitivity, the number of false-positive marks was high; the readers were exposed to 1.3 false-positive marks per film; and several obvious screen-detected cancers that CAD did not mark might have confused test readers. Moreover, the extra work of comparing film mammograms with small CAD-made paper prints might have bee distracting.

The reader changed a false-positive finding to true-negative ten times. That raised the specificity, a "positive" effect, but which is opposite to the intended use of CAD. In countries where lawsuits against radiologists are common, it may be tempting to use CAD as a backup tool in borderline cases.

A more serious flaw is when a reader misses a cancer when CAD does not mark it. This negative effect happened ten times in test readings. Such a possible negative effect of CAD was also studied by Alberdi et al. (2004); their series of 60 patients was enriched with cancers that CAD did not mark (20 of 30 cancers). They found that their group of radiologists reading mammograms without CAD outperformed the group using CAD. The follow-up study by Alberdi et al., carried out at same time as Study II, found that correct CAD output was likely to help in reaching a correct decision, but the incorrect CAD output made it more difficult. In the ethnographic part of their study, readers vocalized their thought processes as they read cases. The implication was that CAD was being used not only as a detection aid but also as a classification or diagnostic aid, which is not what the tool is designed for (Alberdi 2005). Taplin et al. (2006) also suggested that reviewers may have ignored their own findings because CAD did not mark the lesion.

When CAD is used as intended by the manufacturers, only additional findings can occur. Additional false-positive findings reduce specificity, which is a negative effect. Considering the high rate of false-positive prompts, this negative phenomenon took place rather rarely, 25 times. Maybe because of the number of false-positives, readers ignore them as well as the true-positive prompts among them. And of course some false-positive prompts indicate obvious benign lesions or artefacts and therefore do not put strain on readers.

The time between readings without and with CAD was 7 days as a minimum, which raises questions about the memory effect. Nevertheless, there also occurred changes in individual readings contrary to CAD prompts, which speaks against a significant memory effect; Readers lost cancer 11

times, found new cancers 5 times, found new false-positives 5 times, and 15 times correctly changed a finding to true-negative. Perhaps the intra-observer variation explains part of this phenomenon; the threshold of suspicion unconsciously fluctuates from one day to the next. However, individual sensitivities remained at the same level in both rollers. This phenomenon strengthens the assumption that readers ignored most of the prompts by CAD.

The assumption that CAD helps especially inexperienced reader was not verified, as no clinically important differences existed between group performances in readings with and without CAD. Some trends emerged; the negative effect was seen more often among non-screeners, and the screeners found more cancers with CAD aid than did others. The least sensitive readers ignored most of the markings by CAD, but still one of them (a screener) found two additional cancers with CAD aid. Thurfjell et al. (1998) and Balleyguier et al. (2005) found in their studies that CAD helped more novice readers. They had only two readers in their test sets, cancers were screen-detected, and all readings happened in same session. Taylor et al (2004) reported that the readings without and with CAD appeared in different sessions, as in our study, but they had 50 readers, which gives more statistical power than in the other studies. They found no difference in performances with CAD between radiologists and radiographers.

Such studies which lack any prior teaching of CAD have received criticism. Readers had no prior experience with CAD, but they became familiar with CAD while reading films of 100 women in the first roller. After that, they benefited no more from CAD while reading the films of another 100 women on the second roller. It is hard to create a test situation which sufficiently simulates daily practice. Many tests, like ours, are enriched with cancers, so the proportion of true positive marks by CAD is higher than in practice. The cancer cases were selected, and the proportion of difficult cases was high, which also differs from practice. To achieve more statistical power, we would have needed more voluntary readers and also more cancer cases and more normal mammograms. However, our case set (51 cancers/200 women) corresponds with those of other studies similar to ours (Ciatto et al. 2003, Alberdi et al. 2004, Taylor et al. 2004).

The most comprehensive study to date of CAD influence on mammography screening includes data from 332 869 mammograms interpreted without and 24 770 mammograms interpreted with the use of CAD. CAD was implemented in 7 of 43 facilities in the USA during the 4-year study period. Fenton et al. (2007) found that use of CAD (ImageChecker system, R2 Technology) was associated with significantly higher false-positive rates, recall rates, and biopsy rates and with significantly lower overall accuracy in screening mammography than was nonuse. As the increase in sensitivity was nonsignificant and associated strongly with the detection of DCIS (clustered microcalcification), the effect on mortality may have been limited. They estimated that to detect one

additional invasive or in situ cancer, CAD would generate approximately 157 recalls and 15 biopsies, and that nationwide use of CAD would increase the annual costs of screening mammography by 18% ($550 million). The weakness in their study was the wide confidence intervals around estimates of sensitivity and cancer detection rates after implementation of CAD, which is a consequence of a low number of cancers. Because of the rarity of breast cancer, 750 000 mammograms would need to be interpreted to achieve high statistical power (Fenton 2007).

The CAD system (Second Look™ v. 4.01) used in present study was impractical; the machine itself was space-consuming and expensive, films had to be independently fed to the film digitizer which was also time-consuming, and eventually comparing small paper printouts to mammograms caused extra work. These drawbacks are avoided when the CAD system can be included in full-field digital imaging, especially when the software becomes cheaper.

New techniques to improve mass lesion detection have been under development. New features have been designed, for example to measure the degree of sharpness and microlobulation of mass margins for a means to discriminate malignant from benign mass lesions (Varela et al. 2006). Because the first findings of cancers are often increasing densities or temporal changes in parenchymal architecture, a temporal change analysis has been developed. This technique links a suspicious location on the current mammogram with a corresponding location on the prior mammogram. Similarity features measure whether these two regions are comparable in appearance and may prove useful for lesions visible on the prior view as well as for newly developing lesions. Significant ($p = 0.005$) improvement resulted when 465 temporal pairs (238 benign and 227 malignant views) were read without and with the temporal change analysis (Timp et al. 2007). While improving mass lesion detection, these new techniques may lead to a decrease in the false-positive rate, which has been confusingly high. Results in future may favor CAD more than the studies do today.

6.4. Number of readers

In the test group, none of the readers found exactly same cancers or made the same false-positive findings. The sensitivities were low as mentioned earlier, but the specificities were good, over 95%. Accuracies were over 88%, indicating that readers were more likely to find cancers than misinterpret benign conditions as cancer. A trend was that the least-sensitive readers on the whole made fever findings, leading to better specificities. The fact that readers tend to focus on different features aroused curiosity; is it possible to determine an optimal number of readers who find most

cancers? In some countries, a third reader or panel of several readers is used to solve discordant double readings. For this reason, conference consensus readings, where the positive opinion of the majority is considered decisive, were compared to independent readings of a different number of readers. Calculations were performed in two directions: beginning either from the most-sensitive or from the least-sensitive reader to evaluate extreme situations. Starting from the more sensitive reader, the sensitivity stopped increasing after four readers (maximal sensitivity was 74.5% compared to 56.9% for the best single reader). But calculation starting from the other end showed that it was impossible to determine the optimal number, because sensitivity increased until all eight readings were summarized.

Because readers diagnosed different kinds of lesions, almost all of the results improved from the other reader's findings in independent double-reading. Of 28 such pairs 25 pairs differed, and in those three pairs with sensitivity not increasing, the specificity decreased only 0.3-1.3% from the specificity of the better reader. Double readings were simulated independent readings, so the sensitivity was probably higher than it would have been in consensus double readings where one reader may be overruled by another.

The summarized independent readings yielded better sensitivities but inferior specificities than did conference consensus readings. The specificities were still ≥ 90%. Conference consensus readings with only two readers obviously gave the same result as summarized independent readings of two, because instead of true consensus, a positive finding made by one of the two readers made the group decision automatically positive.

The calculations concerning two groups of three readers showed the most interesting results concerning the use of these different methods in practice: with summarized reading it was possible to find 9 and 11 more cancers and 13 and 7 extra false-positives than in conference consensus readings.

In prospective studies by Duijm et al. (2004) and Ciatto et al. (2005), the use of an arbitration panel had a limited negative effect in terms of a reduced cancer detection rate, but the advantages – the reduction of overall referral rates and saved cost per woman screened – were also limited. Duijm et al. (2004) also suggested further diagnostic assessment whenever two independent readers do not reach a consensus, and Ciatto et al. (2005) suggested that new studies with higher recall rates before arbitration may be recommended as cost-effective (Duijm 2004, Ciatto 2005).

As noted earlier, results from this kind of artificial test are not directly generalizable to common practice. Although a strong trend appeared that three readers can find more cancers than two with only a minor decrease in specificity, it is clear that use of this reading method is impractical, because of lack of funds and of mammography screeners. What also has to be kept in

mind is that unnecessary recall causes anxiety and in some cases groundless fear of cancer in otherwise symptomless women.

In Finland, as in many other European countries, a consensus double reading is preferred to an independent reading. During the year 2005, 180 156 women were screened in Finland with 98% specificity, 2.7% were recalled and 4.8 cancers found per 1 000 screening studies (Finnish Cancer registry). In the Netherlands, the recall rates are the world's lowest (<1%), one Dutch study concluded that it could be increased: Recall rates of 1 to 4% were found to be most beneficial in terms of earlier detection of cancers (Otten et al. 2005). In North America, where the recall rates are higher than in Europe, target recall rates were estimated based on over 2 million screening mammograms. Target recall rates of approximately 10% for the first and 6.7% for subsequent mammograms were recommended (Schell et al. 2007). Independent double reading with referral if any reader suggested, it was found to be cost-effective compared to double reading with consensus, in a Dutch study where 500 test cases were read by 26 screening radiologists (Groenewoud et al. 2007).

In Netherlands, two additional readers were used in screening. In the study by Duijm et al (2007), 61 251 screening mammograms were independently read by two radiographers and two radiologists. After standard double-reading involving referral of 905 women with 323 cancers, radiologists reviewed 446 additional images according to the radiographers' suggestions. This review produced an additional 80 referrals, which led to diagnosis of 22 new cancers, and a rise in the recall rate from 1.5% to 1.6%. As 2-year follow-up showed that referral of all 1351 positive readings of those four readers would have led to detection of a total of 362 cancers (a relative increase of 12%) while maintaining a low recall rate. Therefore this study suggested that a referral strategy that includes all radiographer-positive readings should be considered (Duijm 2007).

In the situations where two readers do not easily reach a consensus it might be beneficial to recall. As in present study, longer experience does not guarantee better performance, and a risk always exists that a senior radiologist may overrule a junior in daily practice.

Perhaps in the future, resources could be focused on difficult cases: over 75% of "missed" lesions were found in normal/dense breasts. The possible use of additional readers could focus on these dense breasts, or perhaps only a single reading be done for fatty breasts already selected by the radiographer.

6.5. Breast biopsy

As early detection of breast cancer is essential, so is correct preoperative diagnosis. Even though FNAC has been abandoned in many modern breast centres, it is still widely used in the units referring patients to Breast Surgery Unit of Helsinki University Central Hospital. Many of those units have not introduced CNB in their clinics, or prefer FNAC mainly because it is quicker to perform and in some cases cheaper for the patient. In contrast, a standard practice in the breast-imaging division of The Department of Radiology of HUCH is to use primarily CNB for suspicious breast lesions.

Both the complete and absolute sensitivity of CNB were better than those of FNAC regardless of tumor characteristics. The absolute sensitivity of US-guided FNAC was not as high as in the review by Britton 67% versus 83% otherwise the sensitivities were in agreement with those of other studies (Gordon et al. 1993, Parker et al. 1994, Britton 1999, Westenend et al. 2001, Lieske et al. 2006, Ciatto et al. 2007). Spiculated mass lesions in MG are regarded as highly predictive for malignancy (Thurfjell et al. 2002). This feature was combined with a suspicious finding on ultrasound. Almost all, (99%) of these lesions proved to be invasive carcinomas. Even in these lesions, CNB was superior to FNAC regardless of lesion size.

The method of classifying final outcome including only histologically verified subjects causes verification bias in estimating test accuracy (Houssami et al. 1998). Presenting true sensitivities and specificities of FNAC and CNB in our region would have required the annual statistics of all biopsies taken in all referral units and a minimum of 12 months follow-up. It was impossible to include lesions considered as benign based on findings from the clinical examination, imaging and needle biopsy, because these benign lesions were neither operated on nor were followed up at our unit. In addition, most patients with lesions considered benign are not even referred to our unit. But this situation does not prefer either biopsy method. For these reasons, we do not know the number of false-negatives or true-negatives, so the findings of this retrospective study do not reflect the true sensitivities of FNAC and CNB, and the specificities cannot be calculated. For some measurements, the term positive predictive value would be more precise.

The biopsies were taken and analyzed by several radiologists and pathologists with varying levels of experience. This has certainly influenced the results, such as for FNAC sensitivity, although the sensitivity of FNAC did not differ whether it was taken at HUCH or in other units. On the other hand, the present study setting reflects the situation in many breast surgery centers that have several referral units.

The frequent need for repeat biopsies reduced the advantage of the quick results expected of FNAC. In general, it took from 3 to 7 days to get a result from FNAC. However, in some cases the delay from FNAC to surgery was as long as 100 days because of the long waiting lists for additional biopsies and outpatient clinic and surgical treatment, especially during vacation times. In units with immediate reporting of FNAC and further biopsies readily available, the use of FNAC does not cause such a delay in cancer surgery. Nevertheless, many other clinics also have limited radiological, cytological, and surgical resources and may thus encounter the problems addressed here.

FNAC by itself was less expensive than CNB, but the need for extra work-ups and the costs of unnecessary operations raised the total expenses of FNAC over those of CNB. A comparison of the true costs of different biopsy methods is demanding, and the results may not be generalized due to differing treatment practices and unit costs of services. In addition, the success rate of technical performance and analysis of needle biopsy specimens varies, not only between different units, but perhaps also in the same hospital. As was also impossible to obtain annual statistics concerning all the biopsies taken in our region, the overall costs of the whole FNAC series or CNB series cannot be calculated. Measuring all costs from repeated biopsies or unnecessary sentinel node evacuation was also demanding such as the cost of additional sick leave days, or possible future problems in cases of true cancer.

In a single-center study by Bulgaresi et al. (2006) FNAC had very high positive predictive value and therefore was suggested as a useful test in breast diagnosis, especially in assisting clinical decision-making whether to take additional biopsies or to progress to surgical management. Although FNAC was highly sensitive in detecting abnormality, resulting in Papa 3-5 in 94% of the malignant breast lesions in present study, that is not enough for optimal treatment planning. Before surgeon can discuss about different treatment options with patient, plan expensive and complex sentinel node operations, or evaluate the need of neoadjuvant treatment, it is crucial to know about the invasiveness and the type of the cancer. Had the first biopsy method been CNB instead of FNAC in all cases, the saving would have been more than 24%. The problem of the demanding task of cytology reporting was also apparent; false-positive FNAC findings led to unnecessary axillary surgery in three patients. The need for additional biopsies also means more appointments with radiologists and surgeons, which strains the capacity of the public health service. Multiple procedures and delay at any stage in obtaining a definitive diagnosis of both benign and malignant conditions are undesirable outcomes causing patient discomfort and anxiety.

7. CONCLUSIONS

I: Of breast cancers, 52% were visible in prior mammograms 2 to even 4 years before detection. CAD was more sensitive for lesions in preoperative mammograms, but still found over 60% of lesions in prior mammograms, with specificity being poor on both.

II: As CAD did not significantly improve mammography reading performance among screeners or inexperienced mammography readers, the use of applied version of CAD (Second Look™ v. 4.01) is questionable.

III: The sum of the independent reading of four best-performing readers yielded the best sensitivity. Summarized independent readings were more sensitive than the conference consensus reading; the specificity of summarized readings decreased but still remained 90% or more.

IV: In the breast cancer detection CNB was more sensitive and specific than FNAC. The frequent need for additional work-up raised the total expenses of FNAC over those of CNB and seemed to extend time to treatment. It is therefore recommended to use CNB as the initial needle biopsy method.

8. ACKNOWLEDGEMENTS

This project was carried out at the Helsinki Medical Imaging Center, and I would like to express my gratitude to Juhani Ahovuo, the Chief Executive Officer, for placing excellent research facilities at my disposal.

My deepest gratitude goes to Professor Leena Kivisaari, the initiator of this dissertation project and later my supervisor. Without her encouraging support, I would have interrupted this project many times, but her never-ending enthusiasm and optimism carried me to the finish-line.

Sincere thanks also go to Professor Ritva Vanninen from Kuopio and Docent Tarja Rissanen from Oulu for their constructive criticism. The timetable was tight, and I am grateful that despite their many other time-consuming duties they managed to carefully evaluate this manuscript.

Special thanks go to Docent Tapio Vehmas who carried out a large part of this study; he was "the statistical brain" and an endless source of ideas. I respect his commitment to supervising and I thank him for his patience concerning my inability to understand statistics.

I am grateful to Docent Martti Pamilo, one of the pioneers of mammography screening in Finland, whose professional competence I greatly admire. Martti provided me research material as well as a research environment in Mammography Centre at the beginning of this dissertation project. He also did much with CAD and with reading of the study material.

I have been amazed at how much energy and passion for work, for science and for life can go into one small package: Docent Marjut Leidenius, you are truly an exemplary person! I hope that collaboration with you, as well as with my co-authors Karl von Smitten and Päivi Heikkilä, continues in the future.

My dear family, friends and colleagues, thank you for all those many, many, moments, when you have listened my overwhelming complaining, and sometimes swearing, about my incapabilities in "making science". Believe me; in the darkest moments of disbelief and desperation it has been your support that has helped me to carry through. And of course I need to thank you for just being yourselves - splendid company and magnificent personalieties.

My dear son Jooa, You are more that I ever could imagine, I am grateful for these first 11 months of your life (and I also have to thank you for letting mom sleep so long at night☺). Special thanks go for Jooa's grandparents who have been enormous help by babysitting him.

My personal "IT-help", "mental-coach", best friend, beloved husband, and father of Jooa, thank you for the endless support and love! You know how much you mean to me.

This work has been supported by Helsinki University Research Funds, The Radiological Society of Finland and Per Oscar Klingendahl Foundation.

9. REFERENCES

Al Sarakbi W, Salhab M, Thomas V, Mokbel K. Is preoperative core biopsy accurate in determining the hormone receptor status in women with invasive breast cancer. International seminars in surgical oncology 2005;2:15.

Alberdi E, Povyakalo A, Strigini L, Ayton P. Effects of incorrect computer-aided detection (CAD) output on human decision-making in mammography. Acad Radiol 2004;11:909-918.

Alberdi E, Povyakalo AA, Strigini L, Ayton P, Hartswood M, Procter R, Slack R. Use of computer-aided detection (CAD) tools in screening mammography: a multidisciplinary investigation. Br J Rad 2005;78:31-40.

Altomare V, Guerriero G, Giacomelli L, Battista C, Carino R, Montesano M, Vaccaro D and Rabitti C. Management of nonpalpable breast lesions in a modern functional breast unit. Breast Cancer Res Treat 2005;93:85-89.

American College of Radiology Illustrated breast imaging reporting and data system (BI-RADS), 4th edn. American College of Radiology, Reston, VA.

Anttila A, Koskela J, Hakama M. Programme sensitivity and effectiveness of mammography service screening in Helsinki, Finland. J Med Screen 2002;9:153-158.

Anttinen I, Pamilo M, Soiva M, Roiha M. Double reading of mammography screening films- one radiologist or two? Clin Radiol 1993;48:414-421.

Baez E, Huber A, Vetter M, Hackelöer B-J. Minimal invasive complete excision of benign breast tumors using a three-dimensional ultrasound-guided mammotome vacuum device. Ultrasound in obstetrics and gynecology 2003;21:267-272.

Balleyguier C, Kinkel K, Fermanian J, Malan S, Djen G, Taourel P, Helenon O. Computer-aided detection (CAD) in mammography: Does it help the junior or the senior radiologist? Eur J Rad 2005;54:90-96.

Ballo MS, Sneige N. Can core needle biopsy replace fine-needle aspiration cytology in the diagnosis of palpable breast carcinoma. Cancer 1996;78:773-777.

Barlow WE, Chi C, Carney PA, Taplin SH, D'orsi C, Cutter G, Hendrick RE, Elmore JG. Accuracy of screening mammography interpretation by characteristics of radiologists. J Natl cancer Inst. 2004;96:1840-50.

Beam CA, Layde PM, Sullivan DC. Variability in the interpretation of screening mammograms by US radiologists, findings from national sample. Arch Intern Med 1996; 56:209-213.

Beam CA, Conant EF, Sickles EA. Association of volume and volume-independent factors with accuracy in screening mammogram interpretation. J Natl Cancer Inst 2003;95:282-290.

Berg WA, Krebs TL, Campassi C, et al. Evaluation of 14- and 11-gauge directional, vacuum-assisted biopsy probes and 14-gauge biopsy guns in a breast parenchymal model. Radiology 1997;205:203-208.

Berg WA, Gutierrez L, NessAiver MS, Carter WB, Bhargavan M, Lewis RS, Ioffe OB. Diagnostic accuracy of mammography, clinical examination, US, and MR imaging in preoperative assessment of breast cancer. Radiology 2004;233:830-849.

Bird RE, Wallace TW, Yankaskas BC. Analysis of cancers missed at screening mammography. Radiology 1992.

Birdwell RL, Ikeda DM, O'Shaughnessy KF, Sickles EA. Mammographic characteristics of missed cancers later detected with screening mammography and the potential utility of computer-aided detection. Radiology 2001;219:192-202.

Blanks RG, Wallis MG, Given-Wilson RM. Observer variability in cancer detection during routine repeat (incident) mammographic screening in a study of two versus one view mammography. J Med Screen 1999; 6:152-158.

Blanks RG, Moss SM, McGahan CE, Quinn MJ, Babb PJ. Effect of NHS breast screening programme on mortality from breast cancer in England and Wales, 1990-1998: comparison of observed with predicted mortality. BMJ 2000;321:665-669.

Brem RF, Schoonjans JM. Radiologist detection of microcalcifications with and without computer-aided detection: a comparative study. Clin Radiol 2001;56:150-154.

Brem RF, Hoffmeister JW, Rapelyea JA, Zisman G, Montashemi K, Jindal G, Disimio MP, Rogers SK. Impact of breast density on computer-aided detection for breast cancer. Am J Roentgenol. 2005;184:439-444.

Britton PD. Fine needle aspiration or core biopsy. [Review, 31 refs.] The Breast 1999; 8:1-4.

Britton PD, Flower CD, Freeman AH, Sinntamby R, Warren R, Goddard MJ et al. Changing to core biopsy in an NHS breast screening unit. Clin radiol 1997;52:764-767.

Brown J, Stirling B, Warren R. Mammography screening: an incremental cost effectiveness analysis of double reading versus single reading of mammograms. BMJ 1996;312:809-812.

Bulgaresi P, Cariaggi MP, Bonardi L, Carozzi MF, Confortini M, Galanti L, Maddau C, Matucci M, Rubeca T, Turco P, Ciatto S, Miccinesi G. Analysis of morphologic patterns of fine-needle aspiration of the breast to reduce false-negative results in breast cytology. Cancer (cancer cytopathology) 2005;105:152-157.

Bulgaresi P, Cariaggi P, Ciatto S and Houssami N. Positive predictive value of breast fine needle aspiration cytology (FNAC) in combination with clinical and imaging findings: a series of 2334 subjects with abnormal cytology. Breast Cancer Res Treat 2006;97:319-321.

Chen AM, Haffty BG, Lee CH. Local recurrence of breast cancer after breast conservation therapy in patients examined by means of stereotactic core-needle biopsy. Radiology 2002;225;707-712.

Ciatto S, Del Turco MR, Burke P, Visioli C, Paci E, Zappa M. Comparison of standard double reading and computer-aided detection (CAD) of interval cancers at prior negative screening mammograms: blind review. Br J Cancer 2003;89:1645-1649.

Ciatto S, Ambrogetti D, Risso G, Catarzi S, Morrone D, Mantelli P, Rosselli Del Turco M. The role of arbitration of discordant reports at double reading of screening mammograms. J Med screen 2005;12:125-127.

Ciatto S, Houssami N, Ambrogetti D, Bianchi S, Bonardi R, Brancato B, Catarzi S, Risso G. Accuracy and underestimation of malignancy of breast core needle biopsy: the Florence experience of over 4000 consecutive biopsies. Breast Cancer Res Treat 2007;101:291-297.

Conover WJ. Practical non-parametric statistics. John Wiley & sons. 1980:130-133.

Cornford EJ, Evans AJ, James JJ, Burrell HC, Pinder SE, Wilson AR. The pathological and radiological features of screen-detected breast cancers diagnosed following arbitration of discordant double reading options. Clin Radiol 2005;60:1182-1187.

Cunningham MP. The breast cancer detection demonstration project 25 years later. (Guest editorial) CA Cancer J Clin 1997;47:131-133.

Cupples TE, Cunningham JE, Reynolds JC. Impact of computer-aided detection in a regional screening mammography program. AJR 2005;185:944-950.

Dean PB, Pamilo M. Screening mammography in Finland—1.5 million examinations with 97 percent specificity. Mammography Working Group, Radiological Society of Finland. Acta Oncol 1999;38 suppl 13:47-54.

Destounis SV, DiNitto P, Logan-Young W, Bonaccio E, Zuley ML, Willison KM. Can Computer-aided Detection with double reading of screening mammograms help decrease the false-negative rate? Initial experience. Radiology 2004;232:578-584.

Dinnes J, Moss S, Melia J, Blanks R, Song F, Kleijnen J. Effectiveness and cost-effectiveness of double reading of mammograms in breast cancer screening: findings of a systematic review. Breast 2001;10:455-463.

Duijm LEM, Groenewoud JH, Hendriks JHCL, Koning HJ. Independent double-reading of screening mammograms in the Netherlands: effect of arbitration following reader disagreements. Radiology 2004;231:564-570.

Duijm LE, Groenewoud JH, Fracheboud J, de Koning HJ. Additional double reading of screening mammograms by radiologic technologists: impact on screening performance parameters. J Natl Cancer Inst. 2007;99:1162-1170.

Ellis RL, Meade AA, Mathiason MA, Willison KM, Logan-Young W. Ecaluation of computer-aided detection systems in the detection of small invasive breast carcinoma. Radiology 2007;245:88-94.

Elmore JG, Nakano CY, Koepsell TD, Desnick LM, D'Orsi CJ, Ransoohoff DF. International variation in screening mammography interpretations in community-based programs. J Natl cancer Inst 2003;95:1384-1393.

Esserman L, Cowley H, Eberle C, Kirkpatrick A, Chang S, Berbaum K, Gale A. Improving the accuracy of mammography: volume and outcome relationships. J Natl Cancer Inst 2002;94:369-375.

European guidelines for quality assurance in mammography screening – fourth edition. European commission. Luxembourg. Office for official publications of the European communities, 2006.

European Union Council. Council recommendation of 2 December 2003 on cancer Screening. 2003/878/EC. Bruxelles: European Union Council; 2003.

Evans AJ, Kutt E, Record C, Waller M, Bobrow L, Moss S. Radiological and pathological findings of interval cancers in a multi-center, randomized, controlled trial of mammographic screening in women from age 40-41 years. Clin Radiol 2007;62:348-352.

Fenton JJ, Taplin SH, Carney PA, Abraham L, Sickles EA, D'Orsi C, Berns EA, Cutter G, Hendrick RE, Barlow WE, Elmore JG. Influence of computer-aided detection on performance of screening mammography. N Engl J Med2007;356:1399-1409.

Finnish Cancer Registry. Cancer statistics at www.cancerregistry.fi last updated on September 2007.

Fitzal F, Sporn EP, Draxler W, Mittlbock M, Taucher S, Rudas M, Riedl O, Helbich TH, Jakesz R, Gnant M. Preoperative core needle biopsy does not increase local recurrence rate in breast cancer patients. Breast Cancer Res Treat 2006;97:9-15.

Flobbe K, Bosch AM, Kessels AG, Beets GL, nelemans PJ, von Meyenfeldt MF, van Engelshoven JM. The additional diagnostic value of ultrasonography in the diagnosis of breast cancer. Arch intern Med. 2003;163:1194-1199.

Freer TW, Ulissey MJ. Screening mammography with computer-aided detection: Prospective study of 12,860 patients in a community breast center. Radiology 2001;220:781-786.

Frisell J, Eklund G, Hellström L, Somell A. Analysis of interval breast carcinomas in a randomized screening trial in Stockholm. Breast Cancer Res Treat 1987;9:219-225.

Gennaro G, di Maggio C. Dose comparison between screen/film and full-field digital mammography. Eur Rad 2006;16:2559-2566.

Gordon PB, Goldenberg SL, Chan NHL. Solid breast lesions: diagnosis with US-guided Fine-needle aspiration biopsy. Radiology 1993;198:573-580.

Gram IT, Funkhouser E, Tabar L. The Tabar classification of mammographic parenchymal patterns. Eur J Radiol 1997; 24:131-136.

Groenewoud JH, Otten JDM, Fracheboud J, Draisma G, van Ineveld BM, Holland R, Verbeek ALM, de Koning HJ. Cost-effectiveness of different reading and referral strategies in mammography screening in the Netherlands. Breast Cancer Res Treat 2007;102:211-218.

Gur D, Sumkin JH, Rockette HE, Ganott M, Hakim C, Hardesty L, Poller WR, Shah R, Wallace L. Changes in breast cancer detection and mammography recall rates after the introduction of a computer-aided detection system. J Natl Cancer Inst 2004;96:185-190.

Gönen M, Panageas KS, Larson SM. Statistical issues in analysis of diagnostic imaging experiments with multiple observations per patient. Radiology 2001;221:763-767.

Hakama M, Pukkala E, Heikkilä M, Kallio M. Effectiveness of the public health policy for breast cancer screening: population based cohort study. BMJ 1997;314:864-867.

Harvey JA, Fajardo LL, Innis CA. Previous Mammograms in Patients with Impalpable Breast Carcinoma: Retrospective vs Blinded Interpretation. AJR 1993;161:1167-1172.

Hawass NED. Comparing the sensitivities and specificities of two diagnostic procedures performed on the same group of patients. Br J Radiol 1997;70:360-366.

Helvie MA, Hadjiiski L, Makariou E, Chan HP, Petrick N, Sahiner B, Lo SCB, Freedman M, Adler D, Bailey J, Blane C, Hoff D, Hunt K, Joynt L, Klein K, Paramagul C, Patterson SK and Roubidoux MA. Sensitivity of noncommercial computer-aided detecttion system for mammographic breast cancer detection: pilot clinical trial. Radiology 2004;231:208-214.

Ho WT, Lam PW. Clinical performance of computer-assisted detection (CAD system) in detecting carcinoma in breasts of different densities. Clin Radiol 2003;58:133-136.

Hoffmeister JW, Rogers SK, DeSimio MP, Brem RF. Determining efficacy of mammographic CAD systems. J Dig Imaging 2002;15:198-200.

Houssami N, Irwig L. Likelihood ratios for clinical examination, mammography, ultrasound and fine needle biopsy in women with breast problems. Breast 1998;7:85-89.

Houssami N, Ciatto S, Ambrogetti D, Catarzi S, Risso G, Bonardi R, Irwig L. Florence-Sydney breast biopsy study: Sensitivity of ultrasound-guided versus freehand fine needle biopsy of palpable breast cancer. Breast Cancer Res Treat 2005;89:55-59.

Houssami N, Ciatto S, Ellis I, Ambrogetti D. Underestimation of malignancy of breast core-needle biopsy. Cancer 2007(a);109:487-495.

Houssami N, Ciatto S, Bilous M, Vezzosi V, Biianchi S. Borderline breast core needle histology: predictive value for malignancy in lesions of uncertain malignant potential (B3). Br J Cancer 2007(b);96:1253-1257.

Ikeda DM, Andersson I, Wattsgård C, Janzon L, Linell F. Interval carcinomas in the Malmö mammoGraphic screening trial: radiographic appearance and prognostic considerations. AJR 1992;159:287-294.

Ikeda DM, Birwell RL, O'Shaughnessy KF, Sickles EA, Brenner RJ. Analysis of 172 subtle findings on prior normal mammogram in women with breast cancer detected at follow-up screening. Radiology 2003;226:494-503.

Jackman RJ, Nowels KW, Rodriguez-Soto J, Marzoni Jr FR, Finkelstein SI, Shepard MJ. Stereotactic, automated, large-core needle biopsy of nonpalpable breast lesions: false-negative and histologic underestimation rates after long-term follow-up. Radiology 1999;210:799-805.

Jiang Y, Nishikawa RM, Schmidt RA, Toledano AY, Doi K. Potential of computer-aided diagnosis to reduce variability in radiologists' interpretations of mammograms depicting microcalcifications. Radiology 2001;220:787-794.

Kan L, Olivotto IA, Warren Burhenne LJ, Sickles EA, Coldman AJ. Standardized abnormal interpretation and cancer detection ratios to assess reading volume and reader performance in a breast screening program. Radiology 2000;215:563-567.

Karssemeijer N, Otten JDM, Verbeek ALM, Groenewoud JH, Koning HJ, Hendriks JHLC and Holland R. Computer-aided detection versus independent double reading of masses on mammograms. Radiology 2003; 227:192-200.

Klemi PJ, Parvinen I, Pylkkänen L, Kauhava L, Immonen-Räihä P, Räsänen O, Helenius H. Significant improivement in breast cancer survival through population-based mammography screening. Breast 2003;12:308-313.

Ko JM, Nicholas MJ, Mendel JB, Slanetz PJ. Prospective assessment of computer-aided detection in interpretation of screening mammography. AJR 2006;187:1483-1491.

Koskela AK, Sudah M, Berg MH, Kärjä VJ, Mustonen PK, Kataja V, Vanninen RS. Add-on device for stereotactic core-needle biopsy: How many specimens are needed for reliable diagnosis? Radiology 2005;236:801-809.

Kuhl CK, Schrading S, Bieling HB, Wardelmann E, Leutner CC, Koenig R, Kuhn W, Schild HH. MRI for diagnosis of pure ductal carcinoma in situ: a prospective observational study. Lancet 2007;370:485-92.

Leach MO, Boggis CR, Dixon AK, Easton DF, Eeles RA, Evans DG, Gilbert F, Griebsch I, Hoff RJ, Kessar P, Lakhani SR, Moss SM, Nerurkar A, Padhani AR, Pointon LJ, Thompson D, Warren RR,; MARIBS study group. Screening with magnetic resonance imaging and mammography of a UK population at high familial risk of breast cancer: a prospective multicenter cohort study. Lancet 2005;365:1769-1778.

Leivo T, Salminen T, Sintonen H, Tuominen R, Auerma K, Partanen K, Saari U, Hakama M, Heinonen OP. Incremental cost-effectiveness of double-reading mammograms. Breast Cancer Res treat. 1999;54:261-267.

Liberman L, Kaplan JB. Percutaneous core biopsy of nonpalpable breast lesions: utility and impact on cost of diagnosis [review, 44 refs.]. Breast disease 2001;13:49-57.

Liberman L, Gougoutas CA, Zakowski MF, LaTrenta LR, Abramson AF, Morris EA, Dershaw DD. Calcifications highly suggestive of malignancy. AJR 2001;177:165-172.

Lieske B, Ravichandran D, Wright D. Role of fine-needle aspiration cytology and core biopsy in the preoperative diagnosis of screen-detected breast carcinoma. British journal of cancer 2006;95:62-66.

Logan-Young W, Dawson AE, Wilbur DC, Avila EE, Tomkiewicz ZM, Sheils LA, Laczin JA, Taylor AS. The cost-effectiveness of fine-needle aspiration cytology and 14-gauge core needle biopsy compared with open surgical biopsy in the diagnosis of breast carcinoma. Cancer 1998;82:1867-1873.

Malich A, Marx C, Facius M, Boehm T, Fleck M, Kaiser WA. Tumour detection rate of a new commercially available computer-aided detection system. Eur Radiol 2001; 11:2454-2459.

Malich A, Fischer DR, Facius M, Petrovitch A, Boettcher J, Marx C, Hansch A, Kaiser WA. Effect of breast density on computer-aided detection. J Digit Imaging 2005;18:227-233.

Miller AB, To T, Baines CJ, Wall C. Canadian National Breast Screening Study-2: 13-year results of a randomized trial in women aged 50-59 years. J Natl Cancer Inst. 2000;92:1490-1499.

Miller AB, To T, Baines CJ, Wall C. Canadian National Breast Screening Study-1: breast cancer mortality after 11 to 16 years of follow-up. A randomized trial in women aged 40-49 years. Ann Intern Med 2002;137:305-312.

Moberg K, Bjurstam N, Wilczek B, Rostgard L, Egge E, Muren C. Computed assisted detection of interval breast cancers. Eur J Radiol 2001;39:104-110.

Morton MJ, Whaley DH, Brandt KR, Amrami KK. Screening mammograms: interpretation with computer-aided detection-prospective evaluation. Radiology 2006;239:375-383.

Moss SM, Cuckle H, Evans A, Johns L, Waller M, Bobrow L; Trial management group. Effect of mammographic screening from age 40 years on breast cancer mortality at 10 years' follow-up: a randomised controlled trial. Lancet 2006;368:2053-2060.

Mushlin AI, Kouides RW, Shapiro DE. Estimating the accuracy of screening mammography: a meta-analysis. Am J Prev Med 1998;14:143-153.

Nath ME, Robinson TM, Tobon H, Chough DM, Sumkin JH. Automated large core needle biopsy of surgically removed breast lesions: comparison of samples obtained with 14-16 and 18-gauge needle.Radiology 1995;197:739-742.

Nyström L, Andersson I, Bjurstam N, Frisell J, Nordenskjöld B, Rutqvist LE. Long-term effects of mammography screening: updated overview of the Swedish randomized trials. Lancet 2002;16:909-919.

Gotzsche PC, Olsen O. Is screening for breast cancer with mammography justifiable? Lancet 2000;355:129-134.

Olsen O, Gotzsche PC. Cochrane review on screening for breast cancer with mammography. Lancet 2001;355:129-134.

Olsen AH, Njor SH, Lynge E. Estimating the benefits of mammography screening; The impact of study design. Epidemiology 2007;18:487-492.

Otten JD, Karssemeijer N, Hendriks JH, Groenewoud JH, Fracheboud J, Verbeek AL, Koning HJ, Holland R. Effect of recall rate on earlier screen detection of breast cancers based on the Dutch performance indicators. J Natl Cancer Inst 2005;97:748-754.

Otto SJ, Fracheboud J, Looman CWN, Broeders M, Boer R, Hendriks J, Verbeek A, de Konig H. Initiation of population-based mammography screening in Dutch municipalities and effect on breast cancer mortality: a systematic review. Lancet 2003;361:1411-1417.

Parker SH, Burbank F, Jackman RJ, Aucreman CJ, Gardenosa G, Cink TM, Coscia JL, Eklund GW, Evans III WP, Garver PR, Gramm HF, Haas DK, Jacob KM, Kelly KM, Killebrew LK, Lechner MC, Perlman SJ, Smid AP, Tabar L, Taber FE and Wynn RT. Percutaneous large-core breast biopsy: a multi-institutional study. Radiology 1994;193:359-364.

Parvinen I, Helenius H, Pylkkänen L, Anttila A, Immonen-Räihä P, Kauhava L, Räsänen O, Klemi PJ. Service screening mammography reduces breast cancer mortality among elderly women in Turku. J Med Screen 2006;13:34-40.

Pijnappel RM, van den Donk M, Holland R, Mali W, Peterse J, Hendrikc j, Peeters P. Diagnostic accuracy for different strategies of image-guided breast intervention in cases of nonpalpable breast lesions. Br J Cancer 2004;90:595-600.

Pisano ED, Fajardo LL, Tsimiskas J, Sneige N, Frable WJ, Gatsonis CA, Evans WP, Tocino I, McNeil BJ. Rate of insufficient samples for fine-needle aspiration for nonpalpable breast lesions in a multicenter clinical trial (The radiologic diagnostic oncology group 5 study). Cancer 1998;82:679-688.

Pisano ED, Fajardo LL, Caudry DJ, Sneige N, Frable WJ, Berg WA, Tocino I, Schnitt SJ, Connolly JL, Gatsonis CA, McNeil BJ. Fine-needle aspiration biopsy lesions in a multicenter clinical trial: results from the radiologic diagnostic oncology group V. Radiology 2001;219:785-792.

Pisano Ed, Gatsonis C, Hendrick E, yaffe M, Baum JK, Acharyya S, Conant E, Fajardo LL, Bassett L, D'Orsi C, Jong R and Rebner M. Diagnostic performance of digital versus film mammography for breast-cancer screening. NEJM 2005;353:27:1773-1783.

Plantade R, Hammou JC, Fighiera M, Aubanel D, scotto A,, Gueret S. [Underestimation of breast carcinoma with 11-gauge stereotactically guided directional vacuum-assisted biopsy]. [abstract of review, 117 refs][french]. Journal of radiology 2004;85:391-401.

Saarenmaa I, Salminen T, Geiger U, Holli K, Isola J, Kärkkäinen A, Pakkanen J, Piironen A, Salo A and Hakama M. The visibility of cancer on earlier mammograms in a population-based screening programme. Eur J Cancer 1999;35:1118-1122.

Saarenmaa I, Salminen T, Geiger U, Heikkinen P, Hyvärinen S, Isola J, Kataja V, Kokko M-L, Kokko R, Kumpulainen E, Kärkkäinen A, Pakkanen J, Peltonen P, Piironen A, Salo A, Talviala M-L and Hakama M. The effect of age and density of the breast on the sensitivity of breast cancer diagnostic by mammography and ultrasonography. Breast Cancer Res Treat 2001;67:117-123.

Sauer G, Deissler H, Strunz K, Helms G, Remmel E, Koretz K, Terinde R and Kreienberg R. Ultrasound-guided large-core needle biopsies of breast lesions: analysis of 962 cases to determine the number of samples for reliable tumour classification. Br J Cancer 2005;92:231-235.

Schell MJ, Yankaskas BC, Ballard-Barbash R, Qaqish BJ, Barlow WE, Rosenberg RD, Smith-Bindman R. Evidence-based target recall rates for screening mammography. Radiology 2007;243:681-689.

Shannon J, Douglas-Jones AG, Dallimore NS. Conversion to core biopsy in preoperative diagnosis of breast lesions: is it justified by results. J Clin Pathol 2001;54:762-765.

Shapiro S, Venet W, Strax P, Venet L, Roeser R. 10-to-14-year effect of screening on breast cancer mortality. J Nat Cancer Inst 1982;69:349-355.

Sickles EA, Wolverton DE, Dee KE. Performance parameters for screening and diagnostic mammography: Specialist and general radiologists. Radiology 2002;224:861-869.

Simpson W, Neilson F, Young JR, and the Northern Region Breast Screening Radiology Audit Group. The identification of false negatives in a population of interval cancers: a method for audit of screening mammography. Breast 1995;4:183-188.

Skaane P, Skjennald A. screen-film mammography versus full-field digital mammography with soft-copy reading: randomized trial in a population-based screening-program—the Oslo II Study. Radiology 2004;232:197-204.

Sun W, Ailing L, Abreo F, Turbat-Herrera E, Grafton WD. Comparison of fine-needle aspiration cytology and core biopsy for diagnosis of breast cancer. Diagn. cytopathol. 2001;24:421-425.

Sutela A, Vanninen R, Sudah M, Berg M, Kiviniemi V, Rummukainen J, Kataja V, Kärjä V. Surgical specimen can be replaced by core samples in assessment of ER, PR and HER-2 for invasive breast cancer. Acta Oncol. 2007;Jun 11;1-9 (Epub ahead of print)

Swedish Organised Service Screening Evaluation Group. Reduction in breast cancer mortality from organized service screening with mammography: 1. Further confirmation with extended data. Cancer Epidemiol Biomarkers Prev. 2006;15:45-51.

Swedish Organised Service Screening Evaluation Group. Reduction in breast cancer mortality from organized service screening with mammography: 2. Validation with alternative analytic method. Cancer Epidemiol Biomarkers Prev. 2006;15:52-56.

Swedish Organised Service Screening Evaluation Group: Effect of mammographic service screening on stage at presentation of breast cancers in Sweden. Cancer 2007;109:2205-2212.

Tabar L, Gad A, Holmberg LH, Ljunquist U, Eklund G, Fagerberg C, et al. Reduction in mortality from breast cancer after mass screening with mammography. Lancet 1985;1:829-832.

Tabar L, Vitak B, Chen H-H, Duffy SW, Yen MF, Chiang CF, Crusemo UB, Tot T, Smith RA. The Swedish Two-County Trial twenty years later. Updated mortality results and new insights from long-term follow-up. Radiol Clin North Am 2000;38:625-651.

Tabar L, Yen M, Vitak B, Chen HT, Smith RA, Duffy SW. Mammography service screening and mortality in breast cancer patients: 20-year follow-up before and after introduction of screening. The Lancet 2003;361:1405-1410.

Tabar L, Dean PD. Mammography and breast cancer: the new era. Int J Gynec & Obstet 2003;82:319-326.

Taplin SH, Rutter CM, Elmore JG, Seger D, White D, Brenner JR. Accuracy of screening mammography using single versus independent double interpretation. AJR 2000;174:1257-1262.

Taplin SH, Rutter CM, Lehman CD. Testing the effect of computer-assisted detection on interpretive performance in screening mammography. AJR 2006;187:1475-1482.

Taylor PM, Champness J, Given-Wilson RM, Potts HWW, Johnston K. An evaluation of the impact of computer-based prompts on screen readers' interpretation of mammograms. Br J Radiol 2004;77:21-27.

Thurfjell E, Thurfjell MG, Egge E, Bjurstam N. Sensitivity and specificity of computer-assisted breast cancer detection in mammography screening. Acta Radiol 1998;39:384-388.

Thurfjell EL, Lernevall KA, Taube AAS. Benefit of independent double reading in a population based mammography screening program. Radiology 1994;193:241-244.

Thurfjell MG, Lindgren A, Thurfjell E. Nonpalpable breast cancer: mammographic appearance as predictor of histologic type. Radiology 2002;222:165-170.

Timp S, Varela C, Karssemeijer N. Temporal change analysis for characterization of mass lesions in mammography. Trans Med Imaging 2007;26:945-953.

Varela C, Timp S, Karssemeijer N. Use of border information in the classification of mammographic masses. Phys Med Biol 2006;21:425-441.

Verkooijen HM. Diagnostic accuracy of stereotactic large-core needle breast biopsy for nonpalpable breast disease: results of a multicenter prospective study with 95% surgical confirmation. Int J Cancer 2002;99:853-859.

Warren Burhenne LJ, Wood SA, D'orsi CJ, Feig SA, Kopans DB, O´Shaughnessy KF, Sickles EA, Tabar L, Vyborny CJ and Ctellino RA. Potential contribution of computer-aided detection to the sensitivity of screening mammography. Radiology 2000;215:554-562.

Way LW. In: Current surgical diagnosis and treatment, 10th edition. Edited by L W Way. Conneticut: Appleton & Lange 1994.

Westenend PJ, Sever AR, Beek-man-de Volver HJC, Liem SJ. A comparison of aspiration cytology and core needle biopsy in the evaluation of breast lesions, Cancer (Cancer Cytopathology) 2001;93:146-150.